Bayesian Analysis (2016) **11**, Number 4, pp. 1165–1201

A New Family of Non-Local Priors for Chain Event Graph Model Selection

Rodrigo A. Collazo* and Jim Q. Smith†

Abstract. Chain Event Graphs (CEGs) are a rich and provenly useful class of graphical models. The class contains discrete Bayesian Networks as a special case and is able to depict directly the asymmetric context-specific statements in the model. But bespoke efficient algorithms now need to be developed to search the enormous CEG model space. In different contexts Bayes Factor scored search algorithm using non-local priors (NLPs) has recently proved very successful for searching other huge model spaces. Here we define and explore three different types of NLP that we customise to search CEG spaces. We demonstrate how one of these candidate NLPs provides a framework for search which is both robust and computationally efficient. It also avoids selecting an overfitting model as the standard conjugate methods sometimes do. We illustrate the efficacy of our methods with two examples. First we analyse a previously well-studied 5-year longitudinal study of childhood hospitalisation. The second much larger example selects between competing models of prisoners' radicalisation in British prisons: because of its size an application beyond the scope of earlier Bayes Factor search algorithms.

Keywords: chain event graph, Bayesian model selection, non-local prior, moment prior, discrete Bayesian networks,, asymmetric discrete models, Bayes factor search.

1 Introduction

Graphical models provide a visual framework depicting structural relations in a way easily appreciated by domain experts. Bayesian networks (BNs) (Neapolitan (2004); Cowell et al. (2007); Smith (2010); Korb and Nicholson (2011)) have been a particularly successful example of this class. However, despite its power and flexibility to model a wide range of problems, a BN also has some well-known limitations. Conditional independence statements coded by a BN are necessarily symmetric and must hold for all levels of the conditioning variables. In many domains it has been discovered that in practice this is not a plausible class of hypotheses: different levels of variables can give rise to different types of dependences, even different collections of relevant variables. To build classes of models that can accommodate such assumptions, various non-graphical methods have now been suggested and appended to the BN framework, including context-specific BNs (Boutilier et al. (1996); Poole and Zhang (2003); McAllester et al. (2008)) and object-oriented BNs (Koller and Pfeffer (1997); Bangsø and Wuillemin (2000)).

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, R.A.Collazo@warwick.ac.uk

†Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, J.Q.Smith@warwick.ac.uk

 DOI: 10.1214/15-BA981

However, an alternative way to address this issue is to use a *different* graphical framework from the BN to capture such asymmetric dependences. One such class is the class of Chain Event Graphs (CEGs) (Smith and Anderson (2008); Thwaites et al. (2008)). This contains all discrete context-specific BNs as a special case. CEGs are closely related to probabilistic decision graphs (Bozga and Maler (1999); Jaeger (2004); Jaeger et al. (2006)). The topology of a CEG is based on an event tree and can directly depict level specific asymmetric statements of conditional independences.

Being built from a tree, a CEG typically has a huge number of free parameters. This profusion of models makes the class of CEGs extremely expressive but also very large. Standard model selection methods have nevertheless been successfully employed for models with small number of variables (Freeman and Smith (2011); Barclay et al. (2013); Cowell and Smith (2014)). However, in order to search this massive space when the model hypotheses concern more than just a few variables, it is necessary to a priori specify those models that are most likely to be useful. One property that is widely evoked is to bias the selection towards parsimony. So methods that a priori prefer smaller models during the automated model selection have been found particularly useful. For instance, in the context of BNs various authors, e.g. Pearl (2009), have pointed out that well-fitting sparse graphs tend to identify more stable underlying causal mechanisms. In a recent study of prior and posterior distributions over BN model spaces, Scutari (2013) argued that in practice there is often only weak evidence of any dependence associated with certain levels of the conditioning variable.

The focus of this paper will be the search over the space of CEGs which can also be expressed as context-specific BNs. This enables us to choose priors on hyperparameters of the different component models so that the higher scoring models tend to be the simpler ones. Most applied Bayes Factor (BF) selection techniques – often based on conjugate priors – use local priors; that is, priors that keep the null model's parameter space nested in the alternative model's parameter space. However, recent analyses of BF model selection in other contexts have suggested that the use of such standard methods and prior settings tends to choose models that are not sufficiently parsimonious. In particular, Dawid (1999, 2011) and Johnson and Rossell (2010) have shown that local priors are prone to cause an imbalance in the training rate since the evidential support grows exponentially under a true alternative model but only polynomially under a true null model.

To circumvent this phenomenon, BN selection methods based on non-local priors (NLPs) – albeit for graphs of Gaussian variables – have been successfully developed, see Consonni et al. (2013); Consonni and La Rocca (2011); Altomare et al. (2013). These priors vanish when the parameter space associated with a candidate larger model are nested into the parameter space of a simpler one. This enables the fast identification of the simpler model when it really does drive the data generation process. An NLP embodies beliefs that the data generation process is driven by a parsimonious model within a formal BF methodology. Robustifying the inference in this way has proven especially efficacious for retrieving high-dimensional sparse dependence structures.

In this paper, both to ensure parsimony and stability of selection to the setting of hyperparameters we define three new families of NLPs designed to be applied specifically

to discrete processes defined through trees: the full product NLPs (fp-NLPs), the pairwise product NLPs (pp-NLPs) and the pairwise moment NLPs (pm-NLPs). Although here these methods are developed for CEG models, they can also be directly extended for example to Bayesian cluster analyses.

We will find that a great advantage of a pm-NLP is that it retains the learning rate associated with more standard priors if the data generating process is the complex model whilst scaling up the learning rate when the simple model is true. This enforces parsimony over the model selection in a direct and simple way, keeping computational time and memory costs under control. The empirical results presented here also indicate that a CEG model search using pm-NLPs is more robust than one using a local prior in the sense that model selection is similar for wide intervals of values of nuisance hyperparameters.

The necessity for heuristic algorithms for CEG model selection has already been stressed in Silander and Leong (2013) and Cowell and Smith (2014). When used in conjunction with greedy search algorithms – often necessary when addressing these massive model spaces – we also show here that a pm-NLP (see Section 3) helps to reduce the incidence of some undesirable properties exhibited by standard Dirichlet local priors or product NLPs (fp-NLPs and pp-NLPs).

The present text begins in Section 2 with a brief description of the class of CEGs. In Section 3, we then examine what happens when we apply the standard local priors and product NLPs to the selection of CEGs and present some arguments in favour of pm-NLPs. We also develop a formal framework that enables us to employ pm-NLPs for CEG model search within our modified heuristic approach. To show the efficacy of our method, Section 4 presents some summaries of extensive computational experiments for model selection. The first of these examples uses survey data concerning childhood hospitalisation. The second example models the radicalisation process of a prison population. We conclude the paper with a discussion.

2 Chain Event Graph

2.1 Christchurch Health and Development Study Data Set

To illustrate how the CEG can be used to describe a discrete process, we will first revisit the data set used in Barclay et al. (2013) and Cowell and Smith (2014). Later we will use this survey to explore various features of CEG model selection in this problem. The data we use is a small part of the Christchurch Health and Development Study (CHDS) conducted at the University of Otago, New Zealand; see Fergusson et al. (1986) and Barclay et al. (2013) for more details. This was a 5-year longitudinal study of rates of childhood hospitalization here modelled as a function of three explanatory variables:

- Family social background, a categorical variable differentiating between high and low levels according to educational, socio-economic, ethnic measures and information about the children's birth.

- Family economic status, a categorical variable distinguishing between high and low status with regard to standard of living.
- Family life events, a categorical variable signalising the existence of low (0 to 5 events), moderate (6 to 9 events) or high (10 or more events) number of stressful events faced by a family over the 5 years.

One of the many aims of this study was to assess how these three variables might impact the likelihood of childhood hospitalization (a binary variable). We next describe the semantics for a CEG illustrating this using the CEG model discovered in Barclay et al. (2013). In that study, the hospitalisation of a child – the response (and last) variable – is expressed in terms of the following measured sequence of explanatory variables: social status, economic situation, and life events.

2.2 CEG Modelling

The modelling of a process using a CEG requires three steps: the construction of the event tree $\mathscr{T}$ that supports the process; its transformation into the staged tree; and finally the construction of CEG itself (Smith and Anderson (2008); Thwaites et al. (2008); Smith (2010); Freeman and Smith (2011)). User-friendly introductions to this modelling procedure can also be found in Barclay et al. (2013) and in Cowell and Smith (2014).

Recall that an event tree provides a visual representation of the multiple ways that a process can unfold for each unit. The vertices of the tree symbolise specific situations s encountered by a unit during the process. The outgoing edges represents events that may occur immediately after arriving at each given situation s. Note that a situation s is an intermediate state of a possible final result of the process under analysis. In this sense, the situation s is determined by the successive events along its root-to-s path. The floret $F(s)$ is a star $\mathscr{T}$-subgraph that is rooted at a situation s and includes all emanating edges of s to the possible situations a unit arriving at s might traverse next (Freeman and Smith (2011)). To better understand the parametrisation of a CEG, take a sample $\boldsymbol{y} = \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_R\}$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iL_i})$, and where y_{ij} represents the number of units that arrive at situation s_i and then proceed to its emanating edge j in a event tree.

Figure 1 depicts the process associated with the CHDS data set using an event tree. For example, a child in the initial situation s_0 can unfold into the situation s_1 where her family enjoy good social status. She might then experience a comfortable economic background, situation s_3, or a deprived one, situation s_4. The floret associated with the situation s_1 is presented in bold. The situation s_3 represents the state of a child whose family enjoy both a high social status and prosperous economic conditions. For the purpose of CEG learning and model selection, the data set should then report the number of children that passed along each edge (y_{ij}) in this event tree.

By associating a conditional probability to each edge emanating from a situation s given that a unit is at s, we embellish the event tree into a probability tree. Another

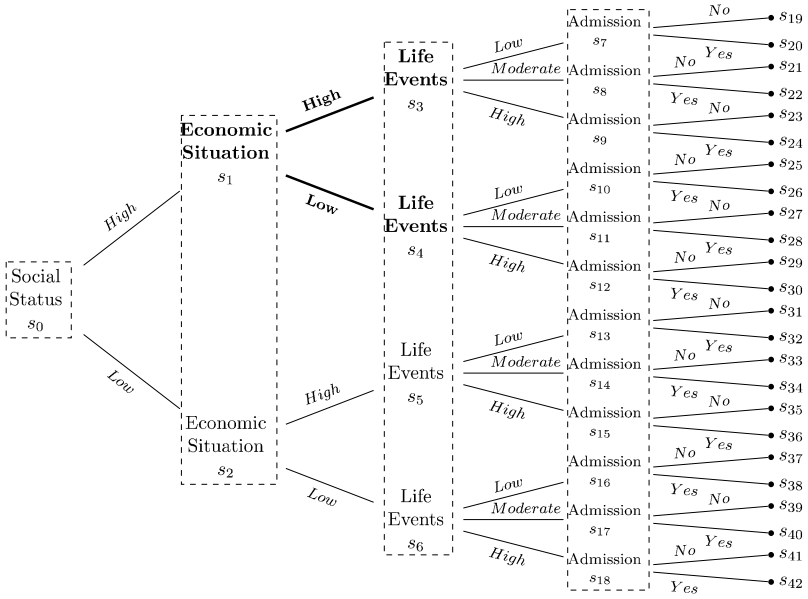


Figure 1: Event Tree associated with the CHDS data set.

construction is a useful basis for depicting a possible evolution. Thus the event tree becomes a staged tree when its situations are coloured. Two situations with the same colour are hypothesised to have the same edge probabilities in the floret they root. To make their association explicit floret edges whose root situations will be assigned the same probabilities are also coloured the same; see Freeman and Smith (2011).

All situations within a given coloured subset (called stage) in the staged tree are said to be in the same position w if they unfold under the same probability law. For a unit arriving at any situation in a particular position the process behind its subsequent evolution will then be identical to those arriving at the other situations in this position. These positions form the vertices of a new graph called a CEG.

The CEG is constructed directly from the staged tree. It simplifies the graph and so expatiates better explanations to domain experts about the hypotheses embodied within the chosen model. Within this construction all leaf nodes are diverted into a single sink node. All situations in the same position are then identified with one another by a single node labelling that position. For the sake of clarity and economy, a position coincident with its stage will be showed in black in the CEG; otherwise, it will keep the colour of its stage in the staged tree.

To introduce the parametrisation, consider a CEG $\mathbb{C}$ which has $M+1$ stages where each stage u_i has L_i emanating edges. Suppose we have a sample $\boldsymbol{x} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_M\}$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iL_i})$, and where x_{ij} represents the number of units that arrive at stage u_i and then proceed to its emanating edge j. Then, associated to each stage u_i

is a probability vector $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iL_i})$, where π_{ij} is the conditional probability of a unit in stage u_i proceeds to take the emanating edge j. Note that $\boldsymbol{x}_i = \sum_{s_j \in u_i} \boldsymbol{y}_{s_j}$ and that the topology of a CEG is completely determined by its stage tree. In fact, the positions are determined once a stage structure is defined as in Figure 2.

Figure 2 shows a possible CEG for the CHDS data set. The positions $w_3 = \{s_3, s_4\}$ and $w_4 = \{s_5\}$ are represented in blue (and in bold) because their corresponding positions s_3, s_4 and s_5 are in the same stage, however their subsequent unfolding is not identical. So, the conditional probabilities associated with variable Life Events are equal given that the variables Social Status and Economic Situation of a family do not simultaneously assume the value "Low". On the other hand, situations in the set $\{s_3, s_4\}$ and the situation s_5 are assigned to different positions because the children of families with low number of stressful Life Events unfold for position w_6 if they are at position $w_5 = \{s_3, s_4\}$ or to position w_7 if they are at position $w_4 = \{s_5\}$. The rest of the positions are also a single stage and are therefore depicted in black. Observe that in contrast to BNs it is very easy and direct to depict asymmetric statements of conditional independence using CEGs.

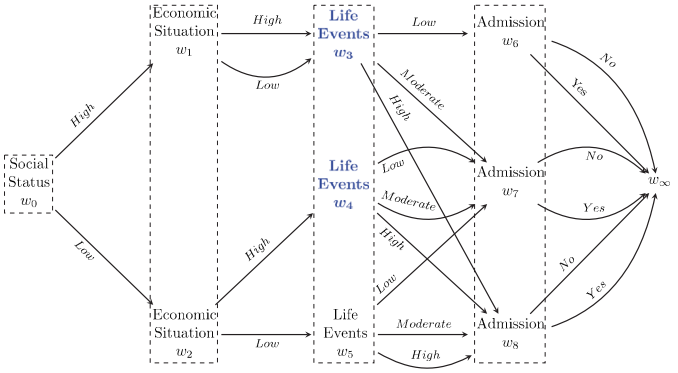


Figure 2: The CEG is associated with the CHDS data set. This figure should be seen in colour for a better understanding.

A triad $\mathbb{C} = (\mathscr{T}, U, \mathscr{P})$ formally characterises a CEG, where $\mathscr{T}$ is an event tree, U is the set of stages, and $\mathscr{P}$ is the adopted probabilistic measure. The pair $\mathbb{G} = (\mathscr{T}, U)$ defines the graphical structure of a CEG $\mathbb{C}$. For the purpose of this paper it is useful to introduce the following property.

Definition 1 (m-Nested Chain Event Graphs)**.** A CEG $\mathbb{C}^+ = (\mathscr{T}, U^+, \mathscr{P})$ is m-nested in any CEG $\mathbb{C} = (\mathscr{T}, U, \mathscr{P})$ if and only if U is a finer partition of U^+ and $|U| - |U^+| = m$. Conventionally Δ is the set of stages of U that are merged in U^+.

2.3 CEG Learning Process

Assume that the $\boldsymbol{\pi}_i$ vectors are mutually independent a priori (floret and path independence condition) and two identical stages in different CEGs in the same probability space have the same prior distribution (staged consistency condition). Then under these two conditions and a complete random sample $\boldsymbol{x}$, Freeman and Smith (2011) proved that each stage in a CEG model space must have Dirichlet distributions a priori and a posteriori. The marginal likelihood under this prior is then given by

$$p(\boldsymbol{x}|\mathbb{G}) = \prod_{i=1}^{M} \frac{\Gamma(\sum_{j=1}^{L_i} \alpha_{ij})}{\Gamma(\sum_{j=1}^{L_i} \alpha_{ij}^*)} \prod_{j=1}^{L_i} \frac{\Gamma(\alpha_{ij}^*)}{\Gamma(\alpha_{ij})}, \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha_{ij}^* = \alpha_{ij} + x_{ij}$ and α_{ij} is the hyperparameter of the Dirichlet prior distribution with regard to emanating edge j from stage u_i.

The hyperparameter $\boldsymbol{\alpha}$ in this prior family plays the role of a phantom sample initialising the CEG learning. Of course, when addressing model selection, it would be impossible to reflect on the massive number of values of possible explanatory hyperparameter vectors and specify them individually. So in practise one common way to sidestep this issue – and one we adopt here – is to fix a hyperparameter $\bar{\alpha}$ and assume a conserving and uniform propagation of this hyperparameter over the event tree. The conserving condition ensures that the total phantom units that emanate from a stage u_i is equal to the total phantom units that arrive at it. Formally, $\bar{\alpha}_i = \sum_{r \in pa(u_i)} \alpha_{i_r j_\star} = \sum_{j=1}^{L_i} \alpha_{ij}$, where $pa(u_i)$ is the set of stages that are parent of u_i and $j_\star$ is the edge that unfolds from a parent stage $u_{i_r} \in pa(u_i)$ to u_i. The uniform assumption implies that the numbers of phantom units that proceed to any two each emanating edges of a stage u_i are identical. This then makes $\alpha_{0j} = \frac{\bar{\alpha}}{L_0}, j = 1, \ldots, L_0$, and $\alpha_{ij} = \frac{\bar{\alpha}_i}{L_i}, i = 1, \ldots, M$, $j = 1, \ldots, L_i$. For instance, take the CEG in Figure 2 and fix $\bar{\alpha} = 6$. So, $\boldsymbol{\alpha}_0 = (3, 3)$, $\bar{\alpha}_1 = \bar{\alpha}_2 = 3$ and $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = (1.5, 1.5)$, where $u_0 = \{w_0\}, u_1 = \{w_1\}$ and $u_2 = \{w_2\}$. Note that there are three edges arriving in stage $u_3 = \{w_3, w_4\}$. Thus, $\bar{\alpha}_3 = 4.5$ and $\boldsymbol{\alpha}_3 = (1.5, 1.5, 1.5)$. The other hyperparameters can be set in a similar way. Henceforth, for any n-dimensional vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{in})$ let $\bar{\gamma}_i = \sum_{j=1}^{n} \gamma_{ij}$.

2.4 Standard CEG Model Selection Using Bayes Factor

Freeman and Smith (2011) developed a framework for implementing a Bayesian agglomerative hierarchical clustering (AHC) algorithm (see, e.g. Heard et al. (2006)) to search over the CEG model space $\boldsymbol{\mathcal{C}}$ for any specific variable order. The AHC algorithm is a greedy search strategy used in conjunction with the log posterior BF. At each iteration, it looks for the MAP model among those 1-nested candidates that result from merging two different stages $u_1, u_2 \in \mathbb{C}$ that have the same number of emanating edges into one stage $u_{1\oplus 2} \in \mathbb{C}^+$ leaving all other stages untouched. By choosing a uniform prior over the model space $\boldsymbol{\mathcal{C}}$ given a variable order, $p(\mathbb{C}) = \frac{1}{|\boldsymbol{\mathcal{C}}|}, \forall \mathbb{C} \in \boldsymbol{\mathcal{C}}$, it was shown that the log-posterior BF (lpBF) between the initial model $\mathbb{C}$ and the candidate model $\mathbb{C}^+$ satisfies:

$$lpBF(\mathbb{C},\mathbb{C}^+) = a(\boldsymbol{\alpha}_1) - a(\boldsymbol{\alpha}_1^*) - b(\boldsymbol{\alpha}_1) + b(\boldsymbol{\alpha}_1^*) + a(\boldsymbol{\alpha}_2) - a(\boldsymbol{\alpha}_2^*) - b(\boldsymbol{\alpha}_2) + b(\boldsymbol{\alpha}_2^*)$$
$$-a(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2) + a(\boldsymbol{\alpha}_1^*+\boldsymbol{\alpha}_2^*) + b(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2) - b(\boldsymbol{\alpha}_1^*+\boldsymbol{\alpha}_2^*), \quad (2)$$

where $a(\boldsymbol{\alpha}_p) = \ln\Gamma(\bar{\alpha}_p)$ and $b(\boldsymbol{\alpha}_p) = \sum_{i=1}^{k_p} \ln\Gamma(\alpha_{pi})$. Note that because the explored model space $\boldsymbol{\mathcal{C}}$ is defined with respect to a particular variable order no two CEGs can be Markov equivalent. So in this sense the setting of the prior is less contentious than it might otherwise be; for such a discussion with respect to BN's see, e.g. Heckerman (1999); Korb and Nicholson (2011).

Barclay et al. (2013) used a BN model search to look for the best variable order in that restricted class. Then to embellish the MAP BN, they employed the AHC algorithm to search for further asymmetric context-specific conditional statements that might be present in the data, using that variable order. One of the highest scoring CEGs is given in Figure 2. Cowell and Smith (2014) then refined this methodology, developing a dynamic programming (DP) algorithm that was able to search a special CEG class called stratified CEG (SCEG) without a pre-defined variable order; see also Silander and Leong (2013).

Sadly this full search method quickly becomes infeasible as the number of explanatory variable increases to an even moderate size. The authors (Silander and Leong (2013); Cowell and Smith (2014)) both recognised that heuristic search strategies would usually be needed when the size of the model space was scaled up. Exploring fast approximations to this approach, Silander and Leong (2013) demonstrated that the AHC algorithm – the method we choose here in our examples – performed better than, for example, methods based on K-mean clustering.

3 Using Non-Local Priors for CEG Model Selection

For CEG model selection, we need to determine when it is better to hold situations apart or merge these into a single stage. The standard BF score can induce rather strange optimal combinations of stages, when the compared stages have very different visit rate ($\bar{\phi}_i$). Theorem 1 below provides us the asymptotic form of $lpBF$ using Dirichlet local priors and makes explicit why difficulties can arise in this context.

Let $\boldsymbol{\phi_i} = (\phi_{i1},\ldots,\phi_{iL_i})$ denote a vector whose element ϕ_{ij} corresponds to the probability of an individual arriving at a stage u_i and taking the emanating edge j of u_i. Then clearly $\phi_{ij} = \bar{\phi}_i * \pi_{ij}$. So each stage u_i can be associated with a random variable $\Phi_i \sim Bernoulli(\bar{\phi}_i)$ that represents whether an individual visits that stage. Analogously each emanating edge j of a stage u_i can be linked to the level of a random variable $\Phi_{ij} \sim Bernoulli(\phi_{ij})$ representing that an individual takes that edge.

Theorem 1. *Take two CEGs $\mathbb{C}$ and $\mathbb{C}^+$ such as $\mathbb{C}^+$ is 1-nested in $\mathbb{C}$. Assume that stages $u_1, u_2 \in \mathbb{C}$ are merged into the stage $u_{1\oplus 2} \in \mathbb{C}^+$. Consider also the true positive conditional probabilities $\boldsymbol{\pi}_1^\dagger$ and $\boldsymbol{\pi}_2^\dagger$ as well as the true positive probabilities $\boldsymbol{\phi}_1^\dagger$ and $\boldsymbol{\phi}_2^\dagger$ associated with stages u_1 and u_2, respectively. If both CEGs have the same prior distribution over the model space $\boldsymbol{\mathcal{C}}$ (see Section 2.4), then as $n \to \infty$*

$$lpBF[\mathbb{C},\mathbb{C}^+] \xrightarrow{a.s.} nB(\boldsymbol{\pi}_1^\dagger,\boldsymbol{\pi}_2^\dagger,\boldsymbol{\phi}_1^\dagger,\boldsymbol{\phi}_2^\dagger) - \frac{L-1}{2}\log(n) + A(\boldsymbol{\phi}_1^\dagger,\boldsymbol{\phi}_2^\dagger,\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2), \quad (3)$$

where A and B are constants that depend on their arguments as given above, and n is the sample size.

Proof. See Appendix A. □

Note that the evidence in favour of any model depends on the sign of the constant B that is analysed in the next two corollaries. As expected, Corollary 1 tells us that there is an imbalance between the learning rates of simple and complex models since the evidence grows logarithmically if the true model is the simple one and linearly otherwise.

Corollary 1. *Take two CEGs $\mathbb{C}$ and $\mathbb{C}^+$ as defined in Theorem 1. If $\boldsymbol{\pi}_1^\dagger = \boldsymbol{\pi}_2^\dagger$, then $B = 0$.*

Proof. This follows directly from equation (29) in the Appendix A. □

Corollary 2 tell us that in any agglomerative search those stages that are more likely to be visited tend to attract stages that are only visited rarely. This is regardless of the generating processes that characterises the conditional probability distributions of these stages.

Corollary 2. *Take two CEGs $\mathbb{C}$ and $\mathbb{C}^+$ as defined in Theorem 1. Consider $\boldsymbol{\phi}_2^\dagger = \kappa \boldsymbol{\phi}_1^\dagger$ where κ is a positive real constant and $\boldsymbol{\pi}_1^\dagger \neq \boldsymbol{\pi}_2^\dagger$. Then, for sufficiently small κ, $B < 0$ regardless of the true conditional probabilities $\boldsymbol{\pi}_1^\dagger$ and $\boldsymbol{\pi}_2^\dagger$.*

Proof. See Appendix B. □

Define the distance between any two stages as given by the distance between their associated expected floret edge probabilistic vectors. According to Corollary 3, massive (or often visited) stages tend to attract to them very light (or less visited) ones no matter how far away these other light stages are in the probabilistic space. Obviously, this is not ideal for highly separated stages to be combined together: they clearly make very different predictions about what will happen to a unit arriving there. Corollary 3 also shows that in contrast, even if other massive stages are very close to each other and so natural to combine, these stages will be less prone to be amalgamated together than in the previous case. Although this is a familiar problem in classical hypothesis testing where statistically different hypotheses might not be significantly different from an interpretative viewpoint, this is nevertheless not a desirable property for Bayesian search algorithms.

Corollary 3. *Take three CEGs $\mathbb{C}$, $\mathbb{C}_1^+$ and $\mathbb{C}_2^+$ where $\mathbb{C}_1^+$ and $\mathbb{C}_2^+$ are 1-nested in $\mathbb{C}$. Assume also that the CEG $\mathbb{C}^\dagger$ is the true model and that this is m-nested in the CEG $\mathbb{C}_1^+$ but is not nested in the CEG $\mathbb{C}_2^+$. If the two stages we combine in CEG $\mathbb{C}$ to form a CEG $\mathbb{C}_2^+$ fulfil the conditions of Corollary 2, then as $n \to \infty$*

$$lpBF[\mathbb{C}, \mathbb{C}_2^+] - lpBF[\mathbb{C}, \mathbb{C}_1^+] \xrightarrow{a.s.} nB_2 + A_2 - A_1 \tag{4}$$

where A_1 is a constant as defined in (3) for SCEGs $\mathbb{C}$ and $\mathbb{C}_1^+$, A_2 and B_2 are the corresponding constants given in (3) for CEGs $\mathbb{C}$ and $\mathbb{C}_2^+$ and where $B_2 < 0$.

Proof. The result follows directly from Corollaries 1 and 2. □

So in this sense the standard BF score can lead to poor model choice when a *pairwise selection* process like the AHC algorithm is used with Dirichlet local priors. The AHC algorithm, which is based on such a sequence of pairwise selection steps, can therefore be sometimes led away from selecting an appropriate model. In fact, this phenomenon is actually exacerbated because of the sequential nature of the AHC algorithm. Once a stage with high true visit rate attracts erroneously other less visited stages, it becomes more massive and therefore more prone to gather incorrectly other smaller stages as the AHC algorithm sequentially agglomerates situations.

The NLP becomes a good option to circumvent this issue. It does this by introducing a formal measure of separation between partitions of the model. This ensures the selection of models not only depends on the probability mass of their partitions but also on the relative distances between their associated probability measures. NLPs therefore provide a promising generic method to more appropriately score CEGs for two main reasons. These priors reduce the imbalance in the learning rate and enforce parsimony in the model selection. They also discourage a greedy model search algorithm from merging two stages spuriously simply because of the probability mass effects discussed above.

To illustrate how we might construct NLPs for CEGs, consider only two variables of the CHDS data set, Social Status and Admission. The corresponding event tree of this process is presented in Figure 3a. Here it is only possible to obtain one of two graphs:

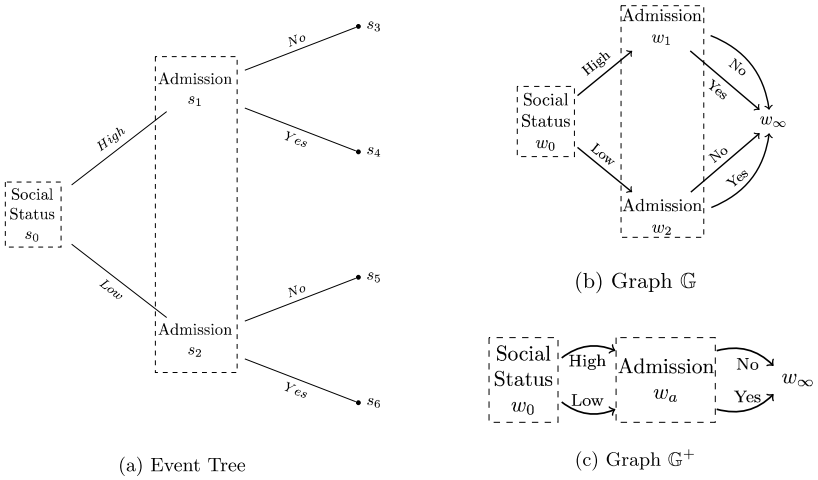


Figure 3: An Event Tree and two possible CEGs that can be modelled using the CHDS data set with *only two* variables, Social Status and Admission.

graph $\mathbb{G}$ with two different stages $u_1 = \{w_1\} = \{s_1\}$ and $u_2 = \{w_2\} = \{s_2\}$ as presented in Figure 3b; or graph $\mathbb{G}^+$ with only one stage $u_a = \{w_a\} = \{s_1, s_2\}$ as presented in Figure 3c, where the stages u_1 and u_2 of $\mathbb{G}$ are merged into a single stage u_a. During the model selection, we need to test whether the stages u_1 and u_2 should be merged or not: $H_0 : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ *vs* $H_1 : \boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$. To do this we construct NLPs that combine the

distance between these two stages $d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ and their probability densities yielded by standard Dirichlet local priors $q_{LP}(\boldsymbol{\pi}_1)$ and $q_{LP}(\boldsymbol{\pi}_2)$. An NLP for the stage u_a of $\mathbb{G}^+$ is equal to its Dirichlet local prior since this stage can not be combined with any other stage: $q_{NLP}(\boldsymbol{\pi}_a|\mathbb{G}^+) = q_{LP}(\boldsymbol{\pi}_a)$ (Figure 4). The NLP density for stages u_1 and u_2 of $\mathbb{G}$ is given by:

$$q_{NLP}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2|\mathbb{G}) = \frac{1}{K} d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho} q_{LP}(\boldsymbol{\pi}_1) q_{LP}(\boldsymbol{\pi}_2), \tag{5}$$

where the proportionality constant $K = E_{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2}[d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho}]$ can be calculated simply using the Dirichlet local priors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ (Figure 5).

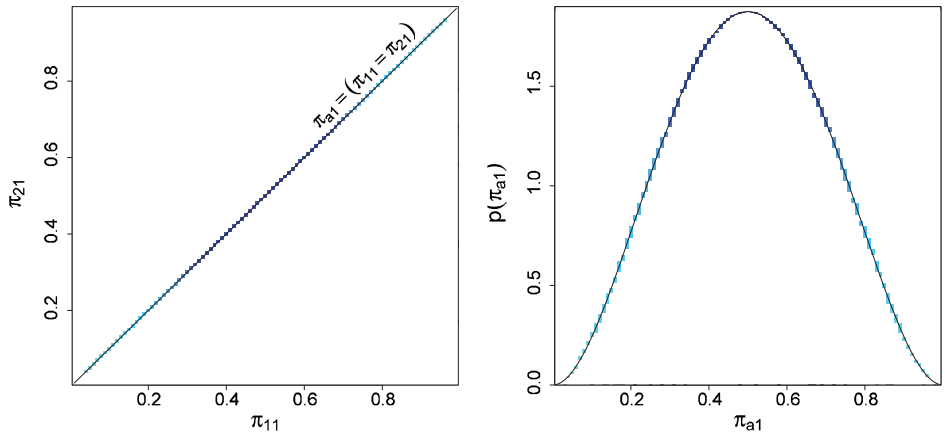
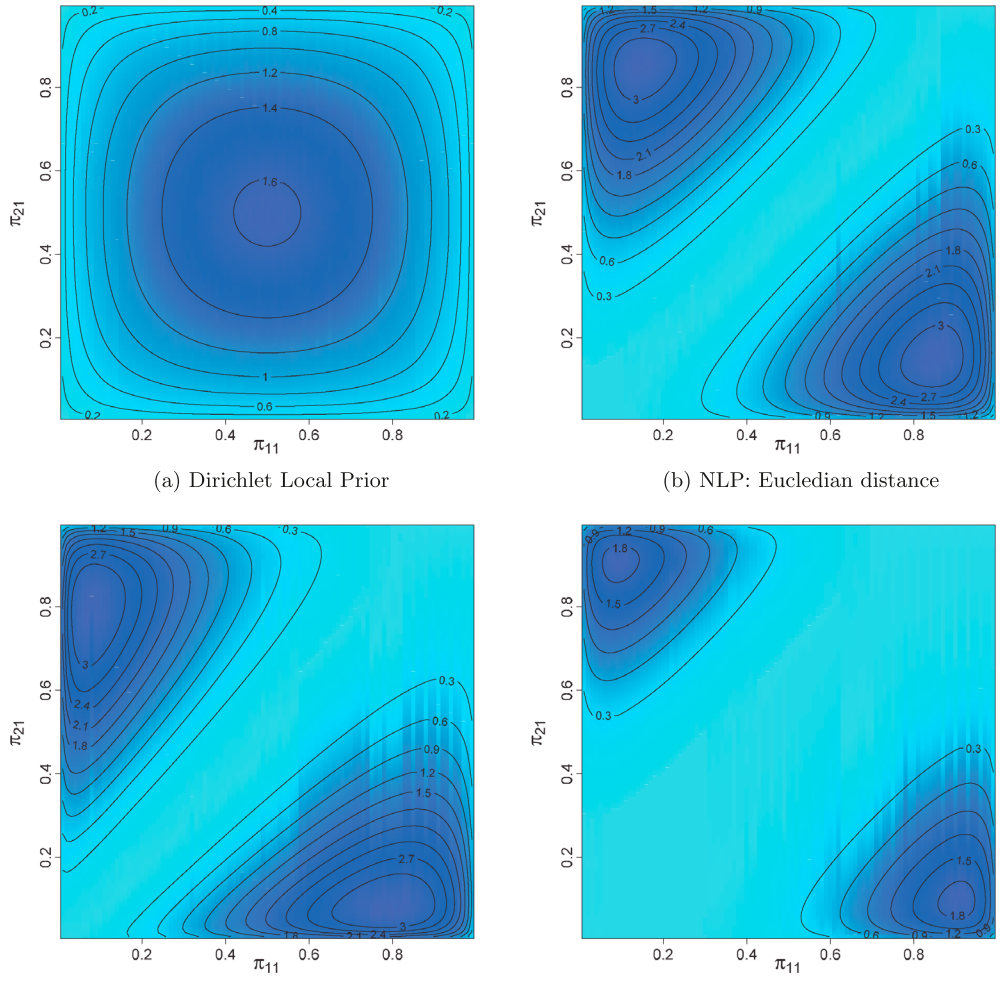


Figure 4: NLP coincident with Dirichlet Local Prior for the only stage associated with the variable Admission in the graph $\mathbb{G}^+$ depicted in Figure 3c where $\boldsymbol{\pi}_a \sim Beta(3, 3)$ and $\bar{\alpha} = 6$. Deeper colour represents higher probability densities.

Note that the NLP for graph $\mathbb{G}$ (see (5)) vanishes when the cell probability vectors associated with the stages u_1 and u_2 are close to one another (Figures 5b, 5c, 5d). Here the probability mass is concentrated a priori in the probability space where the conditional probabilities $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are different. This inhibits the NLP in (5) for the complex model $\mathbb{G}$ from representing the same stage structure ($\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$) which is embedded into the simple model $\mathbb{G}^+$. So, NLPs only allow the parameters corresponding to stages u_1 and u_2 to be identified with each other under the null hypothesis H_0. This contrasts with standard Dirichlet local priors that concentrate the probability mass associated with stages u_1 and u_2 of $\mathbb{G}$ around the probability space where these parameters are equal (Figure 5a). In this sense, local priors do not establish a full partition of the parameter space: the null hypothesis H_0 is nested into the graph $\mathbb{G}$ that should represent only the hypothesis H_1. When using NLPs, these two stages will remain separated or not, based not only on their consistency with the data but also on how far apart these models are, as measured by the distances defined above. Thus as the basis of moderate amount of data, situations tend to be placed in the same stage (graph $\mathbb{G}^+$) unless their edge probabilities are sufficiently different (graph $\mathbb{G}$). We will see at the end of this paper that this enables us to discover models admitting parsimonious explanations as well as good fits to the data.



(a) Dirichlet Local Prior

(b) NLP: Eucledian distance

(c) NLP: Hellinger distance

(d) NLP: Minkowski distance ($\mathcal{L}_4$)

Figure 5: Dirichlet Local Prior and NLPs using different distances for stages associated with the variable Admission in the graph $\mathbb{G}$ depicted in Figure 3b where $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \sim Beta(1.5, 1.5)$ and $\bar{\alpha} = 6$. Deeper colour represents higher contours. Note that the functional forms of different distances are defined in Appendix G.

Remember that we need to elicit a prior joint distribution $p(\boldsymbol{\pi}, \mathbb{G})$ to embed a probabilist map into CEG models. Using Dirichlet local priors and the usual conventions (see, e.g. Heckerman (1999)), the parameter $\boldsymbol{\pi}$ and the graph $\mathbb{G}$ are mutually independent a priori, $p(\boldsymbol{\pi}, \mathbb{G}) = p(\boldsymbol{\pi})p(\mathbb{G})$. This does not happen with NLPs since the prior distribution over the parameter space is conditional on the graph $\mathbb{G}$, $p(\boldsymbol{\pi}, \mathbb{G}) = p(\boldsymbol{\pi}|\mathbb{G})p(\mathbb{G})$. Observe in Figure 5 that given a prior distribution $p(\mathbb{G})$ NLPs reduce the density $p(\boldsymbol{\pi}, \mathbb{G})$ in comparison to local priors only when the distances between the parameters in the corresponding CEGs are close. In contrast, when these distances are substantially dif-

ferent from zero the density indeed increases. In this way, NLPs bias the CEG model selection towards simpler models but only when the data supports them.

Of course, although for simplicity we do not consider this possibility here, we could choose to impose a prior over the model space that further favoured parsimonious models. We note however that although non-uniform priors over the CEG model space reduce the density $p(\boldsymbol{\pi}, \mathbb{G})$ of complex models they do this regardless of the data generation processes. In these cases, the biases in favour of simpler models need to be based on some prior "objective" hypotheses or important prior subjective beliefs over the model space. Despite often being very important in applied studies, these prior distributions are also usually very domain specific. So they are not the focus of this paper.

To extend the previous method of construction of an NLP to the case when there are more than 2 stages (for example, the third level of the CEG in Figure 2), a natural option is to take the product distance between the conditional probability distributions for every pair of stages that can be merged. This family of NLPs is consistent in a sense that their constructions only depend on the characteristics of the particular model associated with that prior. Johnson and Rossell (2012) successfully adopted such a product moment NLP (pMOM-NLP) for Bayesian selection in the context of linear regression. We formally define the fp-NLPs for CEGs below.

In this section, we let $\mathscr{P}_{DLP}$ and $\mathscr{P}_{NLP}$ denote probability measures yielded, respectively, by Dirichlet local priors and NLPs. We also assume that the expectations $E_{\boldsymbol{\pi}}[f(\boldsymbol{\pi})]$ and $E_{\boldsymbol{\pi}*}[f(\boldsymbol{\pi})]$ are calculated, respectively, using the Dirichlet local prior and its corresponding posterior (see Section 2.3) on $\boldsymbol{\pi}$. Finally, in a CEG whose graphical structure is given by $\mathbb{G} = (\mathscr{T}, U)$ we let $\Psi(U)$ denote the collection of pairs of stages (u_i, u_j) in U that can be merged to derive nested CEGs.

To better understand $\Psi(U)$, it is useful to rewrite this as a collection of sets $\{\Psi_k(U)\}_k$, where $\Psi_k(U) = \{u_{r_i}\}_i$ denotes the largest set of stages in U such that the following transitive property holds: for any three stages $u_{r_1}, u_{r_2}, u_{r_3} \in \Psi_k(U)$, if both $(u_{r_1}, u_{r_2}) \in \Psi(U)$ and $(u_{r_2}, u_{r_3}) \in \Psi(U)$, then $(u_{r_1}, u_{r_3}) \in \Psi(U)$. Observe that $\{\Psi_k(U)\}_k$ does not need to be a partition of U, although this property is usually desirable in real-world applications because it simplifies the implementation of model search algorithms. Now we can write

$$\prod_{\{(u_i,u_j)\}\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho} = \prod_{k=1}^{M} \prod_{i=1}^{M_k-1} \prod_{j=2}^{M_k} d(\boldsymbol{\pi}_{r_i}, \boldsymbol{\pi}_{r_j})^{2\rho}, \tag{6}$$

where $M = |\Psi(U)|$ and $M_k = |\Psi_k(U)|$. To illustrate this construction recall the CEG depicted in Figure 2. In this case, the collection $\Psi(U)$ uses the variables that characterise that process. So, in the notation above, we then have that $\Psi(u) = \{\Psi_0 = \{u_0\}, \Psi_1 = \{u_1, u_2\}, \Psi_2 = \{u_3, u_4\}, \Psi_3 = \{u_5, u_6, u_7\}\}$. Here all stages that are associated with the same variable are gathered into the same set $\Psi_k(U)$. For instance, the set Ψ_2 is made up of those stages associated with the variable Life Events. Note that the positions w_3 and w_4 are in the same stage u_3.

Definition 2 (Full Product Non-local Priors for CEGs)**.** The fp-NLPs for a CEG $\mathbb{D} = (\mathscr{T}, U, \mathscr{P}_{NLP})$ where $\mathbb{G} = (\mathscr{T}, U)$ and $\Psi(U) \neq \emptyset$ are given by

$$q_{NLP}(\boldsymbol{\pi}|\mathbb{G}) = \frac{1}{K}\left[\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] q_{DLP}(\boldsymbol{\pi}|\mathbb{G}), \tag{7}$$

where $\rho \in \mathbb{N}^+$ and $K = E_{\boldsymbol{\pi}}[\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}]$ is the normalisation constant. If $\Psi(U)$ is empty then $q_{NLP}(\boldsymbol{\pi}|\mathbb{G}) = q_{DLP}(\boldsymbol{\pi}|\mathbb{G})$.

Assuming random sampling and a non-empty $\Psi(U)$, we can now write the joint distribution of the CEG $\mathbb{D} = (\mathscr{T}, U, \mathscr{P}_{NLP})$ using fp-NLPs as function of the CEG $\mathbb{C} = (\mathscr{T}, U, \mathscr{P}_{DLP})$. Thus,

$$\begin{aligned} p_{NLP}(\boldsymbol{x}, \boldsymbol{\pi}|\mathbb{G}) &= p(\boldsymbol{x}|\boldsymbol{\pi}, \mathbb{G}) q_{NLP}(\boldsymbol{\pi}|\mathbb{G}) \\ &= p(\boldsymbol{x}|\boldsymbol{\pi}, \mathbb{G})\left[\frac{1}{K}\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] q_{DLP}(\boldsymbol{\pi}|\mathbb{G}) \\ &= \left[\frac{1}{K}\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] p_{DLP}(\boldsymbol{x}, \boldsymbol{\pi}|\mathbb{G}). \end{aligned} \tag{8}$$

So, we have that

$$p_{NLP}(\boldsymbol{\pi}|\boldsymbol{x}, \mathbb{G}) = \left[\frac{1}{K^*}\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] p_{DLP}(\boldsymbol{\pi}|\boldsymbol{x}, \mathbb{G}), \tag{9}$$

where $K^* = E_{\boldsymbol{\pi}^*}[\prod_{(u_i,u_j)\in\Psi(U)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}]$ is the normalisation constant. After a little algebra this can be rearranged as

$$p_{NLP}(\boldsymbol{x}|\mathbb{G}) = \frac{K^*}{K} p_{DLP}(\boldsymbol{x}|\mathbb{G}). \tag{10}$$

In this case, the lpBF between two CEGs $\mathbb{D}_1$ and $\mathbb{D}_2$ that have the same prior probability over the model space is given by

$$lpBF(\mathbb{D}_1, \mathbb{D}_2) = lpBF(\mathbb{C}_1, \mathbb{C}_2) + \ln K_1^* - \ln K_2^* - \ln K_1 + \ln K_2, \tag{11}$$

where $\mathbb{C}_1$ and $\mathbb{C}_2$ are the CEGs using Dirichlet local priors that correspond to CEGs $\mathbb{D}_1$ and $\mathbb{D}_2$ using fp-NLPs, respectively. Note that $K = K^* = 1$ if $\Psi(U)$ is empty.

In view of the large size of the CEG space that grows in terms of the Bell number (see Cowell and Smith (2014)), to develop efficient search algorithms it is important to keep calculations as simple as possible, and preferably in closed form. One of the easiest way to do this is to use the Euclidean distance in the formulae above and to set $\rho = 1$. We can also impose a further simplifying condition that $\Psi(U)$ is a partition of the stage set U. But even then in this simple case, for each set $\Psi_k(U) \in \Psi(U)$ of a candidate CEG model we need to calculate a mean of the homogeneous symmetric

polynomial $\prod_{i=1}^{M_k-1}\prod_{j=2}^{M_k} d(\boldsymbol{\pi}_{r_i}, \boldsymbol{\pi}_{r_j})^2$ using the prior and the posterior distributions of the parameters $\boldsymbol{\pi}_i$'s. Note that M_k are often very large since its maximum value depends not only on the number of variables but also on the number of categories that each variable has. The computations can therefore quickly become unmanageable as we scale up the number of variables incorporated into an CEG.

There are also other pitfalls when the fp-NLP is used in conjunction with a greedy search algorithm like the AHC. Using a fp-NLP, Theorem 2 below shows us that the normalisation constant of the posterior distribution of $\boldsymbol{\pi}$ converges to zero with probability 1 if there are at least two stages with the same generating processes. This will happen regardless of whether these stages are under assessment by the model search algorithm. In these cases, Theorem 3 tells us that the marginal posterior probability of such CEG also tends to zero with probability 1. Because of this phenomenon, the fp-NLP is often not a good choice when used in conjunction with a sequential greedy model search even though the method encourages a choice of model with a parsimonious graph.

Let $\boldsymbol{Z}^{(n)} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, where the random variable $\boldsymbol{Z}_s$ registers the events that happen to the sth unit in a process supported by an event tree $\mathcal{T}$. Observe that the event tree $\mathcal{T}$ maps $\boldsymbol{Z}^{(n)} = \boldsymbol{z}^{(n)} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ into a sample $\boldsymbol{x}^{(n)}$ of size n. So, as n increases $\boldsymbol{Z}^{(n)}$ yields a sequence of posterior distributions $p(\boldsymbol{\pi}|\boldsymbol{X}^n, \mathbb{G})$ for the parameter $\boldsymbol{\pi}$. For notational convenience, define a random variable $\boldsymbol{\pi}^*(\boldsymbol{Z}^{(n)}) \sim p(\boldsymbol{\pi}|\boldsymbol{X}^n, \mathbb{G})$ and let $\boldsymbol{\pi}_i^\dagger = (\pi_{i1}^\dagger, \ldots, \pi_{iL_i}^\dagger)$ be the true conditional probability associated with the stage u_i. For clarity, we sometimes write $K^*(\boldsymbol{Z}^{(n)})$ to emphasise that the normalisation constant of a posterior distribution is determined by a sequence $\{\boldsymbol{Z}^{(n)}, n \geq 1\}$.

Lemma 1. *Take the probabilistic parameter π_{ij} associated with the emanating edge j of stage u_i with a positive visiting probability in a CEG $\mathbb{C} = (\mathbb{G}, \mathscr{P}_{DLP})$ and consider $\pi_{ij}^\dagger$ its corresponding true parameter. Then, for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that for all $\epsilon > 0$*

$$\lim_{n\to\infty} P(|\pi_{ij}^*(\boldsymbol{Z}^{(n)}) - \pi_{ij}^\dagger| > \epsilon) = 0. \tag{12}$$

Proof. See Appendix C. □

Theorem 2. *Take a continuous and bounded metric d. In a CEG $\mathbb{C} = (\mathbb{G}, \mathscr{P}_{DLP})$ whose conditional probabilities associated with each edge are strictly positive, for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we then have that as $n \to \infty$*

$$E_{\pi_{ij}^*(\boldsymbol{Z}^{(n)})}\left[\prod_{\substack{(u_i,u_j)\\ \in\Psi(U)}} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] \to \prod_{\substack{(u_i,u_j)\\ \in\Psi(U)}} d(\boldsymbol{\pi}_i^\dagger, \boldsymbol{\pi}_j^\dagger)^{2\rho}. \tag{13}$$

Proof. This follows directly from Lemma 1 and from the continuous mapping theorem (Billingsley (1999)). □

Theorem 3. *Let a CEG $\mathbb{C} = (\mathbb{G}, \mathscr{P}_{DLP})$ have conditional probabilities associated with each edge which are strictly positive. Consider the case when at least two stages in $\mathbb{C}$ have the same true conditional probability according to a continuous and bounded metric d.*

For a CEG $\mathbb{D} = (\mathbb{G}, \mathscr{P}_{NLP})$ *whose probability measure* $\mathscr{P}_{DLP}$ *is generated by a fp-NLP, for almost all sequences* $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ *we then have that as* $n \to \infty$

$$p(\mathbb{D}|\boldsymbol{X}^n, \mathbb{G}) \to 0. \tag{14}$$

Proof. See Appendix D. □

Corollary 4 tells us that when the fp-NLP is used the AHC algorithm can misdirect the search since the normalisation constant of the posterior distribution of $\boldsymbol{\pi}$ may vanish even if the separation between stages does not go to zero in the search neighbourhood. This happens because of the interaction between the definition of fp-NLPs and the data generating process: fp-NLPs are constructed using the product distance between every pair of parameters associated with stages that can be merged ($\Psi(U)$). In contrast, the search neighbourhood defined for the AHC algorithm is only a single pair of stages in $\Psi(U)$. Note that the normalisation constant of the prior distribution of $\boldsymbol{\pi}$ remains unaffected in this case since it is only determined by the phantom sample.

Due to its sequential local strategy, the AHC algorithm can then merge stages that yield the best local score even when this merging is not supported by the data generation process. This situation is further exacerbated because of the combinatorial possibilities that can give rise to circumstances similar to those of Corollary 4. We emphasise that this problem occurs because an fp-NLP is used *in conjunction with* a local search algorithm that for practical reasons we may be forced to adopt: see the comments above. So this is not an issue intrinsically associated with the *form* of an fp-NLP.

Corollary 4. *Take three CEGs* $\mathbb{D}$, $\mathbb{D}_1^+$ *and* $\mathbb{D}_2^+$ *whose probability measures are generated by fp-NLPs using a continuous and bounded metric. Consider that* $\mathbb{D}_1^+$ *merges the stages* u_1 *and* u_2 *of* $\mathbb{D}$, $\mathbb{D}_2^+$ *merges the stages* u_1 *and* u_3 *of* $\mathbb{D}$ *into a new stage* u_a *whose distance to any stage of* $\mathbb{D}_2^+$ *is non-null, and the stages* u_3 *and* u_4 *of* $\mathbb{D}$ *have the same generation process. Assume also that the CEG* $\mathbb{D}^\dagger$ *is the true model that is* 1*-nested in CEG* $\mathbb{D}_1^+$ *but is not nested in CEG* $\mathbb{D}_2^+$*. Then, for almost all sequences* $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ *we have that as* $n \to \infty$

$$\frac{K_1^*(\boldsymbol{Z}^{(n)})}{K_2^*(\boldsymbol{Z}^{(n)})} \to 0, \tag{15}$$

where K_1^* *and* K_2^* *are the normalisation constants with regard to CEGs* $\mathbb{D}_1^+$ *and* $\mathbb{D}_2^+$*, respectively.*

Proof. See Appendix E. □

To sidestep this difficulty, we propose defining NLPs based on pairwise model selection. We note that Consonni and La Rocca (2011) and Altomare et al. (2013) have both used this approach for BN model search. In this framework, the parameters in the contained model have local prior distributions whilst the parameters in the containing model have product NLP distributions. So the choice of prior used in the containing model depends on the contained model. This inconsistency therefore requires a prior specification on the variable order, although in the setting of this paper this order does

not appear to have a significant impact on later inference. The associated ambiguities are extremely small and in practice the method still seems to work well outside this context. Other than this technical nicety, a search method based on these product NLPs enforces parsimony over our model selection whilst allowing us to explore the local properties of our model space. For the CEG family, we call this NLP the pairwise product NLP (pp-NLP).

Given two CEGs whose stage structures U and U^+ are nested ($U^+ \subset U$), recall that the symbol Δ represents the set of stages of U that are merged to obtain U^+ (Definition 1, Section 2.2). Here $\Psi(\Delta)$ denotes the collection of pair of stages (u_i, u_j) in Δ that are gathered in U^+. Analogous to $\Psi(U)$, we can rewrite $\Psi(\Delta)$ as a collection of sets $\{\Psi_k(\Delta)\}_k$. Observe that (16) depends on which pair of CEGs are under analysis whilst (7) is defined in terms of a particular CEG. To illustrate the nature of $\Psi(\Delta)$, take again the stage structure U of the CEG in Figure 2. Consider another CEG whose stage structure U^+ is 3-nested in U in such way that the stages u_1 and u_2 are merged into a stage u_a, and the stages u_5, u_6 and u_7 are combined into a single stage u_b. Then we have that $\Psi(\Delta) = \{\Psi_1 = \{u_1, u_2\}, \Psi_2 = \{u_5, u_6, u_7\}\}$ for the pair of stage structures U and U^+.

Definition 3 (Pairwise Product Non-local Priors for CEGs)**.** To compare the graphical structure $\mathbb{G} = (\mathscr{T}, U)$ with its m-nested graphical structure $\mathbb{G}^+ = (\mathscr{T}, U^+)$, the pp-NLPs for the CEG $\mathbb{D} = (\mathscr{T}, U, \mathscr{P}_{NLP})$ are given by

$$q_{NLP}(\boldsymbol{\pi}|\mathbb{G}) = \frac{1}{K}\left[\prod_{(u_i,u_j)\in\Psi(\Delta)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] q_{DLP}(\boldsymbol{\pi}|\mathbb{G}), \tag{16}$$

where $K = E_{\boldsymbol{\pi}}[\prod_{(u_i,u_j)\in\Psi(\Delta)} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}]$ is the normalisation constant and $\rho = 1, 2, \ldots$

It is easy to see that the complexity of pp-NLPs increases with the number m of nested stages. It can also suffer the same problems as fp-NLPs if the heuristic strategy explores model space neighbourhoods that are smaller than m stages. However, since our goal is only to develop search methodologies when a NLP is used in conjunction with the AHC algorithm, we need to consider only 1-nested CEGs. In this context the pairwise moment NLP (pm-NLP) works well for CEG model search. Comparing (16) and (17), we can see that a pm-NLP is a special case of pp-NLPs when $|\Delta| = 1$.

Definition 4 (Pairwise Moment Non-local Priors for CEGs)**.** To compare the graphical structure $\mathbb{G} = (\mathscr{T}, U)$ and its 1-nested graphical structure $\mathbb{G}^+ = (\mathscr{T}, U^+)$ such as $\Delta = \{u_1, u_2\}$, the pm-NLPs for the CEG $\mathbb{D} = (\mathscr{T}, U, \mathscr{P}_{NLP})$ are given by

$$q_{NLP}(\boldsymbol{\pi}|\mathbb{G}) = \frac{1}{K} d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho} q_{DLP}(\boldsymbol{\pi}|\mathbb{G}), \tag{17}$$

where $K = E_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}[d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho}]$ is the normalisation constant and $\rho = 1, 2, \ldots$

The next corollary shows that a pm-NLP will not exhibit the potential misleading behaviour of the AHC algorithm suffered by product NLPs. The problem is avoided because its normalisation constant only goes to zero with probability 1 if and only if both

merged stages in the contained model have the same generating process. This is because the normalisation constant is defined using exactly the same search neighbourhood as the AHC algorithm - that is, it is a function of densities associated with a single pair of stages. In Corollary 5, $K^* = E_{\boldsymbol{\pi}_1^*,\boldsymbol{\pi}_2^*}[d(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)^{2\rho}]$ is the normalisation constant of the joint posterior distribution of stages u_1 and u_2 when a pm-NLP (Definition 4) is used.

Corollary 5. *Take the CEG D presented in Definition 4such that the metric d is continuous and bounded. Then, for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that*

$$\lim_{n\to\infty} K^*(\boldsymbol{Z}^{(n)}) = 0 \Leftrightarrow d(\boldsymbol{\pi}_1^\dagger, \boldsymbol{\pi}_2^\dagger) = 0. \tag{18}$$

Proof. See Appendix F. □

Now consider a CEG $\mathbb{C} = (\mathscr{T}, U, \mathscr{P}_{DLP})$ and its 1-nested CEG $\mathbb{C}^+ = (\mathscr{T}, U^+, \mathscr{P}_{DLP})$ which aggregates any two stages u_{l1} and u_{l2}. Take the CEG $\mathbb{D} = (\mathscr{T}, U, \mathscr{P}_{NLP})$ whose probability measure is yielded by pm-NLPs. Assuming a uniform prior over the staged structure space, it is straightforward to show that

$$lpBF(\mathbb{D}, \mathbb{C}^+) = \ln \frac{K^*}{K} \frac{p_{DLP}(\boldsymbol{x}|\mathbb{G})}{p_{DLP}(\boldsymbol{x}|\mathbb{G}^+)} \frac{q(\mathbb{G})}{q(\mathbb{G}^+)} = \ln K^* - \ln K + lpBF(\mathbb{C}, \mathbb{C}^+). \tag{19}$$

Pairwise moment NLPs for CEGs can therefore be interpreted as a penalisation over the alternative staged structure U with respect to the distance between the conditional probability distributions of both stages u_1 and u_2. The AHC algorithm can easily be adjusted to incorporate pm-NLPs since we only need to add a term $(\ln K^* - \ln K)$ to the regular $lpBF$ score. So regardless of their minor global inconsistency, the use of a pm-NLP in conjunction with the AHC algorithm is highly computational efficient and also has good local properties.

Define the map G such as $G_y(x) = 1$, if $y = 0$, and $G_y(x) = x$, if $y > 0$, and then the function

$$f(x, y) = \frac{\Gamma(x+y)}{\Gamma(x)} = G_y((x+y-1)\cdot(x+y-2)\cdots x) \tag{20}$$

where x and y are real and natural numbers, respectively. Also let $B(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^n \Gamma(\alpha_j)}{\Gamma(\bar{\alpha})}$ denote the normalisation constant for the Dirichlet distribution parametrised by the vector $\boldsymbol{\alpha}$. The following theorem gives K and K^* of (19) in closed form with regard to the Minkowski distance (see Appendix G). For the corresponding terms with respect to the extension of Hellinger distance to 2ρ-norm spaces, see supplementary material (Collazo and Smith (2015)).

Lemma 2. *Take two random variables $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ which have Dirichlet distributions with parameters $\boldsymbol{\alpha}_1 \in \mathbb{R}_+^L$ and $\boldsymbol{\alpha}_2 \in \mathbb{R}_+^L$, respectively. Define a function $c(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \sum_{j=1}^L (\pi_{1j}^{1/a} - \pi_{2j}^{1/a})^{2\rho}$ where $a > 0$ and $\rho = 1, 2, \ldots$ Then*

$$E[c(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)] = \frac{1}{B(\boldsymbol{\alpha}_1)B(\boldsymbol{\alpha}_2)} \sum_{j=1}^{L} \sum_{h=0}^{2\rho} \left[\binom{2\rho}{h} (-1)^h B(\hat{\boldsymbol{\alpha}}_1^{j,h}) B(\hat{\boldsymbol{\alpha}}_2^{j,h}) \right], \tag{21}$$

where

$$\hat{\alpha}_{1k}^{j,h} = \begin{cases} \alpha_{1k} + \frac{2\rho - h}{a} & \text{if } k = j, \\ \alpha_{1k} & \text{if } k \neq j. \end{cases} \qquad \hat{\alpha}_{2k}^{j,h} = \begin{cases} \alpha_{2k} + \frac{h}{a} & \text{if } k = j, \\ \alpha_{2k} & \text{if } k \neq j. \end{cases}$$

Proof. See Appendix H. □

Theorem 4. *Take the Minkowski distance in a 2τ-norm space ($\tau = 1, 2, \ldots$) to define the pm-NLPs. For the CEG $\mathbb{D}$ presented in Definition 4 whose stages u_1 and u_2 have L emanating edges and $\rho = \tau$, then*

$$K = \sum_{j=1}^{L} \sum_{h=0}^{2\tau} \left[\binom{2\tau}{h} (-1)^h \frac{f(\alpha_{1j}, 2\tau - h) f(\alpha_{2j}, h)}{f(\bar{\alpha}_1, 2\tau - h) f(\bar{\alpha}_2, h)} \right] \tag{22}$$

and

$$K^* = \sum_{j=1}^{L} \sum_{h=0}^{2\tau} \left[\binom{2\tau}{h} (-1)^h \frac{f(\alpha_{1j}^*, 2\tau - h) f(\alpha_{2j}^*, h)}{f(\bar{\alpha}_1^*, 2\tau - h) f(\bar{\alpha}_2^*, h)} \right]. \tag{23}$$

Proof. After some algebra rearrangement this follows directly from Appendix I and Lemma 2 when we set the parameter $a = 1$ and $\rho = \tau$. □

Corollary 6. *Take the Euclidean distance to define the pm-NLPs. For the CEG $\mathbb{D}$ presented in Definition 4 whose stages u_1 and u_2 have L emanating edges and $\rho = 1$, then $K = g(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and $K^* = g(\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*)$ where*

$$g(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \sum_{j=1}^{L} \left[\frac{\gamma_{1j}(\gamma_{1j} + 1)}{\bar{\gamma}_1(\bar{\gamma}_1 + 1)} - 2\frac{\gamma_{1j}\gamma_{2j}}{\bar{\gamma}_1\bar{\gamma}_2} + \frac{\gamma_{2j}(\gamma_{2j} + 1)}{\bar{\gamma}_2(\bar{\gamma}_2 + 1)} \right]. \tag{24}$$

Proof. This follows directly from Theorem 4 when we set the parameter $\tau = 1$. □

Thus we have shown that standard Dirichlet local priors work suboptimally when used in conjunction with the AHC algorithm. This occurs because their corresponding BF scores only take into consideration the probability masses of the stages regardless of their relative location in the probability space. Although it is important not to overstate this problem – conjugate model search is not bad – by introducing a priori a separation measure between stages NLPs tend to perform much better. Their associated BF scores corresponds to the standard local prior BF scores plus a penalisation term as function of the expected distances between stages. However, the use of product NLPs (fp-NLPs and pp-NLPs) is extremely computationally slow. Their penalisation term can also mislead the AHC algorithm since the set of stages used to define them are often bigger than the search neighbourhood of the AHC algorithm (only a pair of stages). In contrast the AHC algorithm using pm-NLPs help us efficiently identify robustly parsimonious models which conjugate or product NLPs cannot. We now illustrate this new selection method.

4 Two Examples for Our Search Method in Action

In this section we compare BF model selection with different non-local and local priors as a function of the hyperparameter $\bar{\alpha}$ using computational simulations based on CHDS data set. These experiments enable us to study how these CEG model selection methods can explain the impact of the explanatory variables appear to have on childhood hospitalisations. We then proceed to analyse the real CHDS data set.

Our second example searches over a much larger space of models. Its hypotheses concern the nature of the radicalisation processes in a prison population. For reasons of confidentiality the data set we used was created through a simulation calibrated to be consistent with publicly available statistics associated with the UK prison population.

Here we use only the simplest possible non-local priors, the quadratic pm-NLPS ($\rho = 1$) associated with Euclidean distance. Although the choice of this metric might superficially look important, at least for the examples we study below the inferences appear robust to this choice. So here we only present and discuss the results using Euclidean distance. A supplementary document (Collazo and Smith (2015)) reports the results using one of the other alternatives – the Hellinger distance. The results using this alternative metric are shown to be remarkably similar to those presented here.

4.1 A CHDS Simulation Study

We based our simulation studies on the CHDS data set assuming, as discussed in Section 2.1, the variable order social status, economic situation, life events and hospital admission. Figure 6 depicts the CEG model we used to generate our simulation experiments. The graphical structure corresponds to a slightly modified version of the MAP CEG found by the DP algorithm under the restriction of that variable order (Cowell and Smith (2014)). Its underlying event tree is presented in Figure 1. The conditional probabilities were assigned based on the real data set. For example, in the CHDS data set 507 individuals enjoy high social status: 53% are in the high economic situation and 47% are in the low economic situation. So, for any unit reaching position w_2 we simulated its next development using a Bernoulli(0.47) random variable.

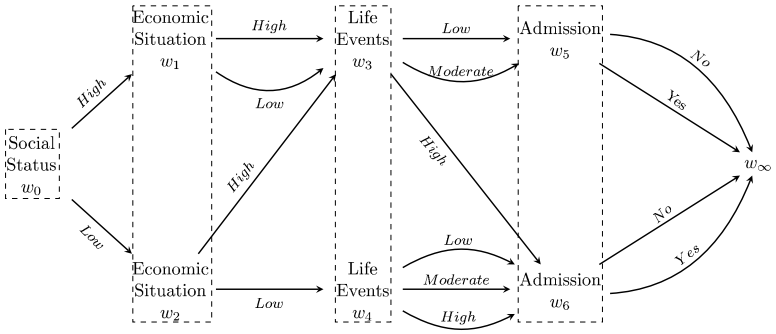


Figure 6: Generating CEG Model for simulation studies with the CHDS data set.

We simulated 100 samples for each sample size (SS) whose range goes from 100 to 5000 by increment of 100. For each sample, the best CEG model was selected by the AHC algorithm for $\bar{\alpha}$-values changing from 1 to 100 by increment of 1 and also for $\bar{\alpha}$-values of 0.1, 0.25, 0.5 and 0.75. We then explored the CEG model space using both Dirichlet local priors and pm-NLPs.

Each CEG chosen was assessed using two criteria: the total number of stages, and the total situational error. The former focus on the topological aspects of the graphical structure. For example, the generating model in Figure 6 has 7 stages. Its objective is to yield a summary of the graphical complexity.

The second criterion checks the overall adequacy of the conditional probabilities associated with each situation of the chosen CEG. This provides us with a diagnostic monitor to assess if the situations in the event tree are merged into stages that indeed represent the data generating model. First, define the empirical mean conditional distributional corresponding to a situation s_j, $\boldsymbol{\mu}(s_j)$, as the mean of the posterior probability distribution of the parameter $\boldsymbol{\pi}_i$ associated with the stage u_i such that $s_j \subset u_i$. Formally,

$$\boldsymbol{\mu}(s_j) = E[\boldsymbol{\pi}_i | \boldsymbol{x}, \mathbb{G}]; \qquad s_j \subset u_i. \tag{25}$$

The situational error $\xi(s_j)$ is the Euclidean distance between the empirical mean conditional distribution and the generating conditional distribution of a situation s_j. Thus

$$\xi(s_j) = \|\boldsymbol{\mu}(s_j) - \boldsymbol{\pi}_i^{\dagger}\|_2; \qquad s_j \subset u_i, \tag{26}$$

where $\boldsymbol{\pi}_i^{\dagger}$ is the conditional probability of the stage u_i in the generating model such that $s_j \subset u_i$. Finally, the total situational error is obtained by the sum of situational errors over the set of situations in the event tree. We therefore have that

$$\xi(\mathcal{T}) = \sum_{j \in \mathcal{T}} \xi(s_j). \tag{27}$$

To analyse the results, average values of each criterion over the 100 data sets for each pair (SS,ESS) were computed. We noted that the corresponding variance is small and does not impact the interpretation of the results presented in Figures 7 and 8. For simplicity, we have depicted below only the outcomes associated with three candidate sample sizes of 300, 900 and 3000. The original study was of 890 children.

Figure 7 shows that pm-NLPs tend to select more parsimonious CEGs than the Dirichlet local priors. Under the assumption that the CEG above is actually the true one we see that the number of stages corresponding to the CEGs chosen by NLPs gets close to the true number (7) of stages over the entire range of $\bar{\alpha}$-values as the sample size increases. In contrast, the CEGs found by local priors are not greatly improved even when the sample size increases from 300 to 3000.

We see in Figure 8 that by selecting simpler graphs NLPs the total of situational errors has reduced. This improves the CEG predictive capabilities. These errors tend to increase for larger values of the parameter $\bar{\alpha}$, particularly for small sample size. The pm-NLPs dominates the local priors consistently for a small sample size and when $\bar{\alpha}$-values

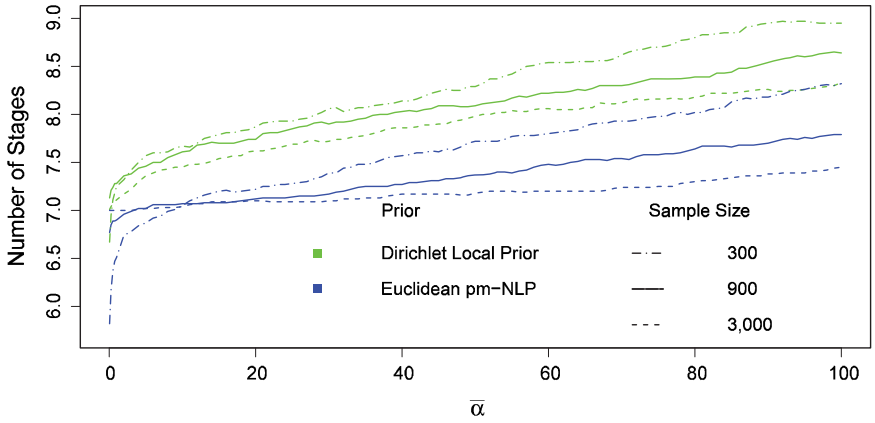


Figure 7: Average of the Number of Stages over the 100 CEGs selected by the AHC algorithm according to the $\bar{\alpha}$-values.

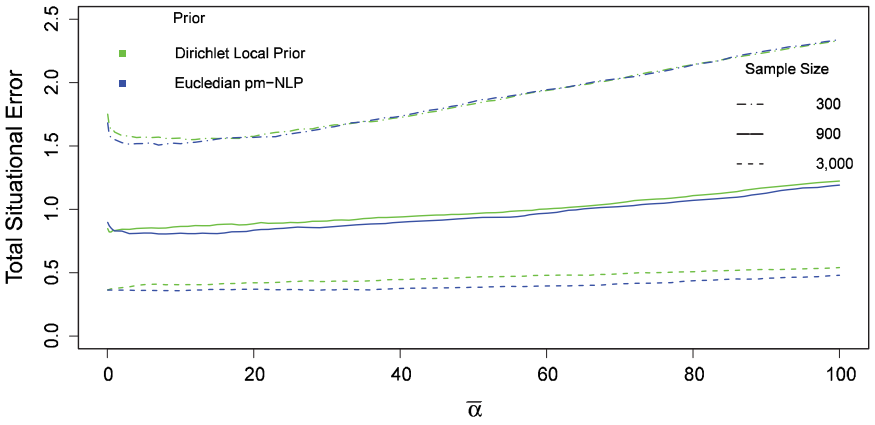


Figure 8: The average of the Total Situational Errors over the 100 CEGs selected by the AHC algorithm according to $\bar{\alpha}$-values.

are not large, and in medium and large sample sizes independently of the $\bar{\alpha}$-values. The best results appear to be concentrated around $\bar{\alpha}$-values from 1 to 20 regardless of the sample size.

The pm-NLPs appear more robust with regard to the hyperparameter $\bar{\alpha}$. They tend to pick more plausible models for values from 1 to 20 of this hyperparameter regardless of the sample size. Observe that in this range the number of stages tend to be quite stable around the true number (7) and the total situational errors are minimised. On the other hand, local priors appear to give rise to substantially different inferences for different values in this parameter range. In this case although it is true that larger values of this hyperparameter give more consistency in terms of the number of stages,

these values imply larger total situational errors. They also represent very strong prior information about the various margins of individual variables: a hypothesis which would usually be a strange one to impose in many practical scenarios.

Lastly, we analyse the influence of very small $\bar{\alpha}$-values (less than 1) on the results. For a sample size of 300, although LPs tend to choose better CEGs than NLPs with regard to the number of stages, these CEGs do not optimise the total situational errors that are indeed slightly greater than those corresponding to CEGs selected by NLPs. Using the medium-size samples (900), LPs lead to more complex CEGs than the true one with respect to the number of stages whilst NLPs tend to select simpler ones, but the number of stages in both cases are the same distance from the true number (7). Here the LPs have barely smaller total situational errors than NLPs. NLPs clearly dominate the local priors in both criteria when the sample size is equal to 3000.

Overall very small $\bar{\alpha}$-values are not recommended since they yield very unstable results using local and non-local priors. They are also inclined to find CEGs with larger total situational errors. In the case of pm-NLPs, these small $\bar{\alpha}$-values tend to select sparser CEG than the true one, having a strong regularization effect over the graphical structure. However, the good modelling practise of calibrating a priori the predictive consequences of such prior settings would usually not encourage the choice of such values.

As expected on the basis of our theoretical results, for this example NLPs tend to be more stable and to select sparser – simpler to explain – graphs especially when compared with conventional methods. The results also indicate that NLPs are more prone to find CEGs that have a slightly better predictive capabilities for all reasonable settings of the hyperparameter $\bar{\alpha}$.

4.2 A New Analysis of the CHDS Data Set

We now compare the performance of our methods using pm-NLPs and Dirichlet local priors in a real analysis of the CHDS data set when the data generating process is assumed unknown. Figure 9 shows how the staged structures change as the parameter $\bar{\alpha}$ increases when we look over the CEG model space using the AHC algorithm under the constraint of the variable order used previously.

Figure 9 enables us to compare the sensitivity of CEG model selection using local and using non-local priors as function of the hyperparameter setting. Note that increasing the stability of the model selection for wider range of $\bar{\alpha}$-values makes the result less dependent on this hyperparameter. The interpretation of the conditional independence statements embedded into the selected CEG then become more reliable since the choice of the CEG is unlikely to change dramaticaly with small perturbation in the $\bar{\alpha}$-values. In fact, it can be seen from Figure 9 that local priors induce more robust results for $\bar{\alpha} \geq 8$, while Euclidean pm-NLPs are quite stable for $\bar{\alpha} \leq 23$. Note that the NLPs provide even more consistent outcomes of the search with regard to small and medium $\bar{\alpha}$-values, i.e. they are more robust to the setting of this hyperparameter than the local priors. Recall from Section 4.1 that better results tend to be obtained by setting $1 \leq \bar{\alpha} \leq 20$.

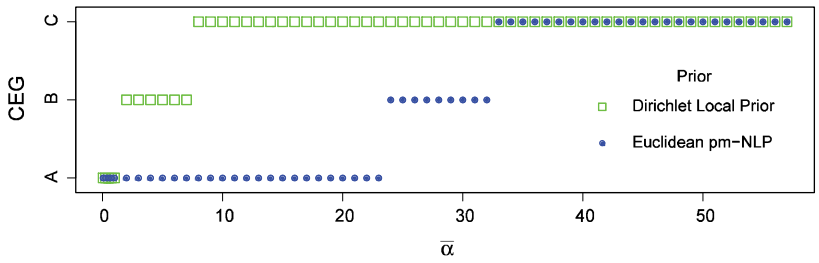
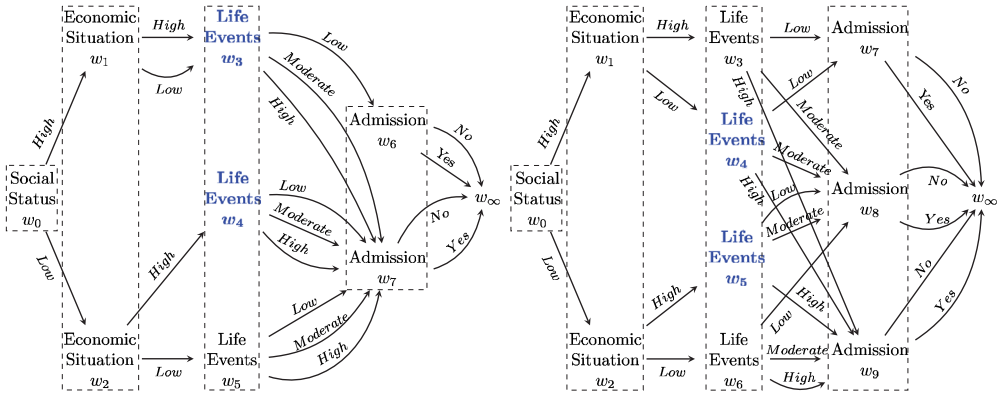


Figure 9: CEG Model Selection for CHDS data set using the AHC algorithm.

Next observe that NLPs tend to select sparser graphs. The CEG A (Figure 10a) has 7 stages, the CEG B (Figure 2) has 8 stages, and the CEG C (Figure 10b) has 9 stages. The AHC algorithm using local priors points to the CEG C whilst the use of pm-NLPs indicates the CEG A. The CEG C is 1-nested and 2-nested in the CEGs B and A, respectively. In fact the qualitative interpretations and the probability measures do not differ very much, although the more parsimonious graphs (e.g. CEG A) give somewhat more transparent and intuitive explanations of the process.



(a) CEG A (b) CEG C

Figure 10: CEGs A and C selected by the AHC algorithm.

Thus observe that the CEG B is identical to the CEG C except that the variable life events has two stages (u_3, u_4) and three positions (w_3, w_4, w_5) in the CEG B, and three stages (u_3, u_4, u_5) and four positions (w_3, w_4, w_5, w_6) in the CEG C. As highlighted in red (Table 1), only the conditional probabilities associated with these positions have changed, and then only very slightly. Furthermore, although these CEGs differ, their causal hypotheses associated with childhood hospitalisation are in fact identical: the hospital admissions are partitioned into the same three groups of patients in both.

Highlighting only the substantial differences implied by the data set, the CEG A brings new and much simplified hypotheses about how hospital admissions relate to

Stage	Conditional Probability Vector	SCEG A ($\bar{\alpha}=3$)	SCEG B ($\bar{\alpha}=6$)	SCEG C ($\bar{\alpha}=12$)
u_0	$(p(S=h), p(S=l))$	(0.57,0.43)	(0.57,0.43)	(0.57,0.43)
u_1	$(p(E=h), p(E=l))$	(0.47,0.53)	(0.47,0.53)	(0.47,0.53)
u_2	$(p(E=h), p(E=l))$	(0.12,0.88)	(0.12,0.88)	(0.13,0.87)
u_3	$(p(L=l), p(L=m), p(L=h))$	(0.46,0.34,0.20)	(0.46,0.34,0.20)	(0.43,0.33,0.24)
u_4	$(p(L=l), p(L=m), p(L=h))$	(0.22,0.31,0.47)	(0.22,0.31,0.47)	(0.5,0.36,0.14)
$-/u_5$	$(p(L=l), p(L=m), p(L=h))$	–	–	(0.22,0.31,0.47)
u_5/u_6	$(p(A=n), p(A=y))$	(0.91,0.09)	(0.91,0.09)	(0.91,0.09)
u_6/u_7	$(p(A=n), p(A=y))$	(0.77,0.23)	(0.82,0.18)	(0.82,0.18)
u_7/u_8	$(p(A=n), p(A=y))$	–	(0.73,0.27)	(0.73,0.27)

Legend: S, Social background; E, Economic situation; L, Life events; A, hospital Admission
l, Low; m, Moderate; h, High; n, No; y, Yes
$\cdot$ / $\cdot$ – SCEGs A & B / SCEG C

Table 1: Conditional probability table for SCEGs found by the AHC algorithm.

the covariates. It proposes the existence of only two distinct risk groups of hospital admission. The CEGs B and C segment the higher risk individuals in the CEG A (position w_7) into two groups (positions w_7 and w_8). Note that the differences in the probability of hospital admission between these two groups (Table 1, in blue) are small. In other words, both groups continue to identify a higher risk population in comparison with individuals who experience a low number of life events and have higher social status.

4.3 The Radicalisation of a Prison Population

Introduction

Our second CEG search was conducted over a much larger class of hypotheses this time about the nature of the process of radicalisation within prisons. The results we give here are a small part of an ongoing study to be reported more fully in a later paper. Our main focus here is to develop methods to identify groups of individuals who are most likely to engage in specific criminal organisation in British prisons. As we will show, this example is very challenging because the classes of each variable are remarkably unbalanced and the percentage of radical prisoners – those units of special interest – is tiny. Furthermore, if expressed in terms of a BN (see Figure 11) any plausible generating model would need to be a highly context-specific: generic BN model selection methods could therefore not be expected to work well. A more flexible family such as the CEG class really does need to be used.

For the purposes of this illustration we have restricted our analyses to consider only six explanatory variables. These have been chosen because they are often hypothesised as playing a key role in the process of radicalisation. These are:

- Gender, a binary variable distinguishing between male (M) and female (F);

- Religion, a nominal variable with three categories: Rel, religious prisoner; NRel, non-religious prisoner; and NRec, not recorded;
- Age, an ordinal variable with three categories: A1, age < 30; A2, $30 \leq$ age < 40; and A3, age ≥ 40;
- Offence, a nominal variable with five categories: VAP, violence against person; RBT, robbery, burglary or theft; D, drug; SO, sexual offence; and O, others;
- Nationality, a binary variable differentiating between British citizens (B) and foreigners (F);
- Network, an ordinal variable differentiating groups of prisoners according to their social interactions with well-known members of the target criminal organisation. It has three categories: I, intense; F, frequent; and S, sporadic.

Because of the sensitive nature of data in this field, we have based this example on a data set some of whose variables have been simulated. However, we have chosen simulations that are calibrated to real figures and real hypotheses in the public domain concerning the British prison population (Ministry of Justice (2013)). So the simulations plausibly parallel the likely current scenario. The generating model used was based on an initially elicited BN depicted in Figure 11. The real data set enables us to naively estimate the joint distributions for the first five explanatory variables. These are presented in black in this figure. Note that several variables have sparse cell counts: for example, Gender (F, 5%), Religion (NRec, 2%) and Nationality (F, 10%).

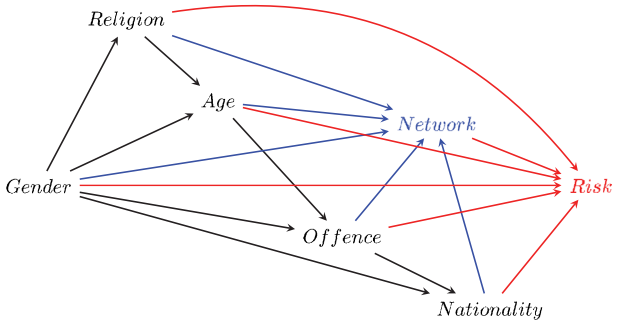


Figure 11: Generating Model for Radicalisation Example.

No data was publicly available for the explanatory variable Network and the response variable Radicalisation. So in this study we instead construct a probability model over certain developments based on expert judgements (Cuthbertson (2004); Jordan and Horsburgh (2006); Hannah et al. (2008); Neumann (2010); Silke (2011); Rowe (2014)). To perform the necessary data simulation, we needed to specify the conditional distribution of variable Network given the first five explanatory variables. Here we assumed that there are only four different social interaction mechanisms. The response variable – introduced last – distinguishes between individuals at high or low risk of radicalisation.

Being the last variable to be sampled for each prisoner, this has 540 conditioning partitions. In this environment risk assessments are generally coarse. So based on the expert judgements cited above, these partitions are clustered into only three different radicalisation classes of risk where the highest risk prisoners come from only six partitions. Note that from a technical viewpoint these plausible hypotheses introduce several prior context-specific conditional assessments into our model.

The radicalisation risk of the whole prison population is hypothesised to be small in line with the expert judgement and academic literature (Cuthbertson (2004); Jordan and Horsburgh (2006); Hannah et al. (2008); Neumann (2010); Silke (2011)). Here this is set at around 0.7% of the total population. Based on the premises discussed above, we then simulated 100 complete data sets. Each of these has 85000 individuals, approximating the recent yearly totals of the British prison population. Assuming our fixed generating model is true we will now investigate the efficacy of various CEG search methods to identify those prisoners most likely to be radicalised in each of these data sets.

CEG Model Searches

Assume that our optimal model is consistent with a variable sequence Gender, Religion, Age, Offence, Nationality, Network and Radicalisation. This simplifies the search space and matches the goals of this work. The CEG model search was performed using a setting of the hyperparameter $\bar{\alpha} = 5$ since this corresponds to the maximum number of categories taken by a variable in the problem. Note that this choice also implies a plausibly large variance over the prior marginal distribution of each variables. Observe that our previous results (Section 4.1) suggests that the selection of a hyperparameter in this region will provide robust results: this was confirmed numerically in additional exploratory studies within this example.

The scale of this problem requires us to use a heuristic algorithm like AHC since the SCEG space contains more than 10^{1105} SCEG models even given the chosen variable order. Here full model search strategies such as ones using Dynamic Programming will obviously be infeasible.

As expected the results in Table 2 indicate that the AHC algorithm in conjunction with Euclidean pm-NLPs was prone to select more parsimonious and user-friendly models than those obtained using standard local priors especially for stages near the leaves of the corresponding event tree. NLPs also ensured that the AHC algorithm selected models with a number of stages associated with the variables Network and Radicalisation closer to the generating model than those achieved using the Dirichlet local priors.

The use of pm-NLPs enabled the AHC algorithm to find CEG models that clearly better represented the simulated generating process of radicalisation. For example, Euclidean pm-NLPs classified the highest risk population spuriously in only 29 data sets whilst local priors had problems with 39 data sets. So local priors misclassified some of the highest risk individuals in more than 34% of the data sets than Euclidean pm-NLPs. These misclassifications using local priors and pm-NLPs were associated with the highest risk groups whose sample sizes were less than 25 and whose sample proportions

Variable Level	Number of Stages using DLP	pm-NLP	Number of Generating Stages	Maximum Number of Stages
Gender	1	1	1	1
Religion	2	2	≤ 2	2
Age	4.8	4.1	≤ 6	6
Offence	6	5.9	≤ 6	18
Nationality	7.4	5.4	≤ 10	90
Network	10.2	7.2	4	180
Radicalisation	7.6	5.6	3	540

Table 2: Average of the Numbers of Stages in Radicalisation CEGs selected by the AHC algorithm using Dirichelet Local Priors and Euclidean pm-NLPs.

of radical prisoners were concentrated around 12%. Furthermore inference using local priors struggled to identify the risk level for a high risk group of 209 individuals where the sample proportion of radical prisoners was 24%.

There were only three levels of risk of radicalisation in the generating model. So for the sake of simplicity the stages that were found by the AHC algorithm were amalgamated in Table 3 according to their corresponding radicalisation risk in five categories. We matched the risks greater than 25%, between 1% and 7% and less than 1% as corresponding to the risk of 30%, 3% and 0.1% in the generating model, respectively.

		Dirichlet Local Prior – Errors					Euclidean pm-NLP – Errors					Number of
CEG Risk(%)		≥ 25	$(15,25]$	$(7,15]$	$(1,7]$	≤ 1	≥ 25	$(15,25]$	$(7,15]$	$(1,7]$	≤ 1	Prisoners
Generating Model Risk	30	–8.9	2.5	3.1	3.0	0.3	–5.5	0	1.6	3.6	0.3	699
	3	16.4	0.2	111	–887	759	19.4	0	57	–844	768	119×10^2
	0.1	0.9	0	3.5	359	–363	1.1	0	1.4	373	–375	724×10^2

Table 3: Average Number of misclassified prisoners over the 100 CEGs selected by the AHC algorithm according to their risk of radicalisation in the Generating Model.

Although local and non-local priors yield broadly equivalent estimates for the lower two levels of radicalisation risk, Dirichlet local priors lost track of 9 of the highly hazardous individuals on average whilst pm-NLPs only lost about 6. This means an improvement of 33% in favour of pm-NLPs. Note also that local priors unlike the pm-NLPs tend to introduce a stage at risk level between 15% and 25%. If we merged the three higher levels of radicalisation risk into one category, we would lose 3 high risk individuals on average regardless of the type of prior used. However, in this case local priors would include 50 more medium risk individuals (3%) in the high category. This would correspond to almost 70% more prisoners that as a result of the analysis would be spuriously identified as a danger to the public.

Although the model used here is rather naive and our results are not perfect, this larger example does nevertheless demonstrate the promise of pm-NLPs used in conjunc-

tion with a greedy search of CEG models when applied to much larger scale asymmetric populations like the one above.

5 Conclusions

The massive CEG model space usually require a heuristic strategy to perform CEG model selections efficiently. We argued here that the product NLPs can inappropriately bias the AHC algorithm since they are defined using a larger set of stages than one used to establish the search neighbourhood (a pair of stages) of the AHC algorithm. However, the product NLP might still be feasible and effective if used in conjunction with a different family of greedy search algorithm. One such example is the weighted MAX-SAT algorithm (Cussens (2008)) where the search neighbourhood can be defined in ways which are neither pairwise nor sequential. These families of algorithms are therefore more compatible with such priors. Preliminary investigation into these methods looks promising and will be reported later. The theorems in Section 3 suggest that exploring stochastic algorithms for use with NLPs looks like another interesting research problem – both for CEGs and other models. Certainly for CEGs this has so far not been attempted to our knowledge. But note that these algorithms tend to use local moves which will suffer from similar problems to those described above unless their construction is carefully designed. A further generalisation of our model selection framework using greedy search algorithms in conjunction with NLPs to more general model classes also looks promising.

Another extension that has not been addressed in this study is to combine the dynamic programming approach with some heuristic strategies for the full product NLPs. The elicitation of product NLPs in closed form is more difficult and their computational implementation is challenging. These drawbacks are accentuated in discrete high-dimensional graphs. Whilst this paper establishes the theoretical background for this purpose, it is necessary to advance algorithmic structures that make good use of computational time and memory in order to scale up to medium size problems. Such algorithms would be a useful diagnostic for assessing the results provided by the AHC algorithm. They could also be used to examine in more detail the robustness of Bayesian CEG model selection to the $\bar{\alpha}$-values.

Most recently dynamic versions of DCEGs have been developed (Barclay et al. (2015)). The corresponding model spaces are then order of magnitude greater than those discussed here. However some initial studies have shown that methods of NLP adopted to this different environment also appear to work well. Results on the search of such spaces will again be reported in a later paper.

Appendix A: Proof of Theorem 1

Using the fact that $\ln\Gamma(z) = (z-0.5) * \ln(z) - z + 0.5 * \ln(2\pi) + O(1)$ as $z \to \infty$ (Abramowitz and Stegun (1972)), we can rewrite (2) as follows:

$$lpBF(\mathbb{C},\mathbb{C}^+) = \sum_{i=1}^{L_1} \alpha^*_{1i} \ln\left(\frac{\alpha^*_{1i}}{\alpha^*_{1i}+\alpha^*_{2i}}\frac{\bar\alpha^*_1+\bar\alpha^*_2}{\bar\alpha^*_1}\right) + \sum_{i=1}^{L_1} \alpha^*_{2i} \ln\left(\frac{\alpha^*_{2i}}{\alpha^*_{1i}+\alpha^*_{2i}}\frac{\bar\alpha^*_1+\bar\alpha^*_2}{\bar\alpha^*_2}\right)$$
$$+\frac{1}{2}\ln\left(\frac{\bar\alpha^*_1\bar\alpha^*_2}{\bar\alpha^*_1+\bar\alpha^*_2}\right) - \frac{1}{2}\sum_{i=1}^{L_1}\ln\left(\frac{\alpha^*_{1i}\alpha^*_{2i}}{\alpha^*_{1i}+\alpha^*_{2i}}\right) + A(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2) + O(1). \quad (28)$$

Using the Strong Law of Large Numbers and the continuous mapping theorem (Billingsley (1999)), we obtain that as $n \to \infty$

$$\begin{aligned} &lpBF(\mathbb{C},\mathbb{C}^+) \\ &\xrightarrow{a.s.} n\left\{\sum_{i=1}^{L_1}\phi^\dagger_{1i}\ln\left[\pi^\dagger_{1i}\left(\frac{\bar\phi^\dagger_1+\bar\phi^\dagger_2}{\pi^\dagger_{1i}\bar\phi^\dagger_1+\pi^\dagger_{2i}\bar\phi^\dagger_2}\right)\right] + \sum_{i=1}^{L_1}\phi^\dagger_{2i}\ln\left[\pi^\dagger_{2i}\left(\frac{\bar\phi^\dagger_1+\bar\phi^\dagger_2}{\pi^\dagger_{1i}\bar\phi^\dagger_1+\pi^\dagger_{2i}\bar\phi^\dagger_2}\right)\right]\right\} \\ &\quad - \frac{L-1}{2}\ln(n) - \frac{1}{2}\ln\left(\frac{1}{\bar\phi^\dagger_1}+\frac{1}{\bar\phi^\dagger_2}\right) + \frac{1}{2}\sum_{i=1}^{L_1}\ln\left(\frac{1}{\phi^\dagger_{1i}}+\frac{1}{\phi^\dagger_{2i}}\right) + A(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2). \end{aligned} \quad (29)$$

Appendix B: Proof of Corollary 2

Assume $D_{KL}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,)$ as the Kullback–Leibler divergence between the discrete probability distributions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Using $\bar\phi^\dagger_2 = \kappa\bar\phi^\dagger_1$ and $\ln(1+z) = z + O(z^2)$ as $z \to 0$, we can rewrite B as follows:

$$\begin{aligned} B &= \bar\phi^\dagger_1\left\{(\kappa+1)\ln(\kappa+1) - \sum_{i=1}^{L_1}\left[\pi^\dagger_{1i}\ln\left(1+\kappa\frac{\pi^\dagger_{2i}}{\pi^\dagger_{1i}}\right) + \kappa\pi^\dagger_{2i}\ln\left(\kappa+\frac{\pi^\dagger_{1i}}{\pi^\dagger_{2i}}\right)\right]\right\} \\ &= \bar\phi^\dagger_1\left[O(\kappa^2) - \kappa\sum_{i=1}^{L_1}\pi^\dagger_{2i}\ln\left(\kappa+\frac{\pi^\dagger_{1i}}{\pi^\dagger_{2i}}\right)\right] \\ &\le \bar\phi^\dagger_1\left[O(\kappa^2) - \kappa\sum_{i=1}^{L_1}\pi^\dagger_{2i}\ln\frac{\pi^\dagger_{1i}}{\pi^\dagger_{2i}}\right] = \bar\phi^\dagger_1\left[O(\kappa^2) - \kappa D_{KL}(\boldsymbol{\pi}^\dagger_2,\boldsymbol{\pi}^\dagger_1)\right]. \end{aligned} \quad (30)$$

Note that the inequality holds because κ is strictly positive. The result follows since the Kullback–Leibler divergence is always non-negative and is equal to 0 if and only if $\boldsymbol{\pi}^\dagger_1 = \boldsymbol{\pi}^\dagger_2$.

Appendix C: Proof of Lemma 1

Using the Strong Law of Large Numbers, it is easy to see that as $n \to \infty$

$$E[\pi^*_{ij}(\boldsymbol{Z}^{(n)})] = \frac{\alpha^*_{ij}}{\bar\alpha^*_i} \xrightarrow{a.s.} \pi^\dagger_{ij}, \quad (31)$$

and

$$Var[\pi^*_{ij}(\boldsymbol{Z}^{(n)})] = \frac{\alpha^*_{ij}(1-\frac{\alpha^*_{ij}}{\bar\alpha^*_i})}{\bar\alpha^*_i(1+\bar\alpha^*_i)} \xrightarrow{a.s.} 0. \quad (32)$$

It follows that

$$E_{\pi^*_{ij}(\boldsymbol{Z}^{(n)})}[(\pi_{ij}-\pi^\dagger_{ij})^2] = Var[\pi^*_{ij}(\boldsymbol{Z}^{(n)})] + (E[\pi^*_{ij}(\boldsymbol{Z}^{(n)})]-\pi^\dagger_{ij})^2 \xrightarrow{a.s.} 0. \tag{33}$$

Since $\pi^*_{ij}(\boldsymbol{Z}^{(n)})$ converges in quadratic means to the true value of the parameter π_{ij} for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$, it also converges in probability to the true value of the parameter π_{ij} for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$. Note that this result also follows directly from Doob's Theorem (see, e.g. Schervish (1996), Section 7.4.1, or DasGupta (2008), Section 20.7).

Appendix D: Proof of Theorem 3

From (10), we have that

$$p(\mathbb{D}|\boldsymbol{x}^{(n)}, \mathbb{G}) = \frac{K^*}{K} p(\mathbb{C}|\boldsymbol{x}^{(n)}, \mathbb{G}). \tag{34}$$

As $0 \leq p(\mathbb{C}|\boldsymbol{x}^{(n)}, \mathbb{G}) \leq 1$ and K is a constant that depends on the hyperparameter $\bar{\alpha}$, we can conclude that

$$\lim_{n\to\infty} K^*(\boldsymbol{z}^{(n)}) = 0 \Rightarrow \lim_{n\to\infty} p(\mathbb{D}|\boldsymbol{x}^{(n)}, \mathbb{G}) = 0. \tag{35}$$

Note now that there are at least two stages u_a and u_b in $\mathbb{C}$ that have the same true conditional probability. So, $d(\pi^\dagger_a, \pi^\dagger_b) = 0$ for some pair of stages u_a and u_b in $\mathbb{C}$. Recall that

$$K^*(\boldsymbol{z}^{(n)}) = E_{\pi^*_{ij}(\boldsymbol{z}^{(n)})}\left[\prod_{\substack{(u_i,u_j)\\ \in\Psi(U)}} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)^{2\rho}\right] \tag{36}$$

Theorem 2 then implies that (35) is always satisfied for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$.

Appendix E: Proof of Corollary 4

Since $\mathbb{D}^\dagger$ is 1-nested into $\mathbb{D}^+_1$, Theorem 2 implies that for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that

$$\lim_{n\to\infty} K^*_1(\boldsymbol{Z}^{(n)}) = 0. \tag{37}$$

On the other hand, $\mathbb{D}^+_2$ does not have stages with equal true conditional probability distribution by construction. Therefore, Theorem 2 also implies that for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that

$$\lim_{n\to\infty} K^*_2(\boldsymbol{Z}^{(n)}) = c \neq 0. \tag{38}$$

The result then follows directly from (37) and (38).

Appendix F: Proof of Corollary 5

From Lemma 1 and from the continuous mapping theorem (Billingsley (1999)), for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that as $n \to \infty$

$$E_{\boldsymbol{\pi}_1^*(\boldsymbol{Z}^{(n)}),\boldsymbol{\pi}_2^*(\boldsymbol{Z}^{(n)})}\left[d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho}\right] \to d(\boldsymbol{\pi}_1^\dagger, \boldsymbol{\pi}_2^\dagger)^{2\rho}. \tag{39}$$

Recall that

$$K^*(\boldsymbol{Z}^{(n)}) = E_{\boldsymbol{\pi}_1^*(\boldsymbol{Z}^{(n)}),\boldsymbol{\pi}_2^*(\boldsymbol{Z}^{(n)})}\left[d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho}\right]. \tag{40}$$

If the necessary condition (18) is true, for almost all sequences $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots)$ we have that as $n \to \infty$

$$K^*(\boldsymbol{Z}^{(n)}) \to 0. \tag{41}$$

Equations (39) and (40) then imply that $d(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)^{2\rho} = 0$.

Assuming $d(\boldsymbol{\pi}_1^\dagger, \boldsymbol{\pi}_2^\dagger) = 0$, the sufficiency follows again directly from (39) and (40).

Appendix G: Minkowski and Hellinger Distances

The Minkowski distance corresponds to a generalisation of the Euclidean distance to the τ-norm space ($\tau = 1, 2, \ldots$). For two points $S = (s_1, \ldots, s_n) \in \mathbb{R}^n$ and $T = (t_1, \ldots, t_n) \in \mathbb{R}^n$, it is given by

$$d(S, T) = \|S - T\|_\tau = \left(\sum_{i=1}^{n} |s_i - t_i|^\tau\right)^{\frac{1}{\tau}}, \tag{42}$$

where $\|\,\boldsymbol{.}\,\|_\tau$ is the τ-norm; see Kruskal (1964) for more details. Note that we have the Euclidean distance when $\tau = 2$.

For two discrete probability distributions $S = (s_1, \ldots, s_n) \in \mathbb{R}^n$ and $T = (t_1, \ldots, t_n) \in \mathbb{R}^n$, the Hellinger distance (Rao (1995)) is defined by

$$d(S, T) = \|\sqrt{S} - \sqrt{T}\|_2 = \left(\sum_{i=1}^{n} (\sqrt{s_i} - \sqrt{t_i})^2\right)^{\frac{1}{2}}. \tag{43}$$

This can be extend to the 2τ-norm space ($\tau = 1, 2, \ldots$) using the formula

$$d(S, T) = \|\sqrt[2\tau]{S} - \sqrt[2\tau]{T}\|_{2\tau} = \left(\sum_{i=1}^{n} (\sqrt[2\tau]{s_i} - \sqrt[2\tau]{t_i})^{2\tau}\right)^{\frac{1}{2\tau}}. \tag{44}$$

Appendix H: Proof of Lemma 2

Expanding the function $c(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ by means of the binomial theorem, we then have that

$$E[c(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)] = \int_0^1 \sum_{j=1}^{L} (\pi_{1j}^{1/a} - \pi_{2j}^{1/a})^{2\rho} \frac{1}{B(\boldsymbol{\alpha}_1)B(\boldsymbol{\alpha}_2)} \prod_{k=1}^{L} \pi_{1k}^{\alpha_{1k}-1} \pi_{2k}^{\alpha_{2k}-1} \mathrm{d}\boldsymbol{\pi}_1 \mathrm{d}\boldsymbol{\pi}_2$$

$$= \frac{1}{B(\boldsymbol{\alpha}_1)B(\boldsymbol{\alpha}_2)} \sum_{j=1}^{L} \int_0^1 \sum_{h=0}^{2\rho} \binom{2\rho}{h} (-1)^h \pi_{1j}^{\frac{2\rho-h}{a}} \pi_{2j}^{\frac{h}{a}} \prod_{k=1}^{L} \pi_{1k}^{\alpha_{1k}-1} \pi_{2k}^{\alpha_{2k}-1} \mathrm{d}\boldsymbol{\pi}_1 \mathrm{d}\boldsymbol{\pi}_2$$

$$= \frac{1}{B(\boldsymbol{\alpha}_1)B(\boldsymbol{\alpha}_2)} \sum_{j=1}^{L} \sum_{h=0}^{2\rho} I_j^h, \tag{45}$$

where

$$I_j^h = \int_0^1 \binom{2\rho}{h} (-1)^h \pi_{1j}^{\frac{2\rho-h}{a}} \pi_{2j}^{\frac{h}{a}} \prod_{k=1}^{L} \pi_{1k}^{\alpha_{1k}-1} \pi_{2k}^{\alpha_{2k}-1} \mathrm{d}\boldsymbol{\pi}_1 \mathrm{d}\boldsymbol{\pi}_2. \tag{46}$$

Let $\hat{\boldsymbol{\alpha}}_1^{j,h}$ such as $\hat{\alpha}_{1k}^{j,h} = \alpha_{1k} + \frac{2\rho-h}{a}$, if $k = j$, and $\hat{\alpha}_{1k}^{j,h} = \alpha_{1k}$, if $k \neq j$. Take also $\hat{\boldsymbol{\alpha}}_2^{j,h}$ such as $\hat{\alpha}_{2k}^{j,h} = \alpha_{2k} + \frac{h}{a}$, if $k = j$, and $\hat{\alpha}_{2k}^{j,h} = \alpha_{2k}$, if $k \neq j$. Then,

$$I_j^h = \binom{2\rho}{h} (-1)^h \int_0^1 \pi_{1j}^{\alpha_{1j}+\frac{2\rho-h}{a}-1} \pi_{2j}^{\alpha_{2j}+\frac{h}{a}-1} \prod_{\substack{k=1 \\ k\neq j}}^{L} \pi_{1k}^{\alpha_{1k}-1} \pi_{2k}^{\alpha_{2k}-1} \mathrm{d}\boldsymbol{\pi}_1 \mathrm{d}\boldsymbol{\pi}_2$$

$$= \binom{2\rho}{h} (-1)^h B(\hat{\boldsymbol{\alpha}}_1^{j,h}) B(\hat{\boldsymbol{\alpha}}_2^{j,h}). \tag{47}$$

Appendix I: Normalisation Constant $B(\alpha)$

If $\nu \in \mathbb{R}_+$ and $z \in \mathbb{N}_+$, than $\Gamma(\nu + z) = \Gamma(\nu) \prod_{i=0}^{z-1} (\nu + i)$ (Abramowitz and Stegun (1972)). Now take $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \boldsymbol{a}$, where $\boldsymbol{\alpha} \in \mathbb{R}_+^n$ and $\boldsymbol{a} \in \mathbb{N}^n$. After using the previous factorisation property of gamma function and organising the products in a convenient way, we obtain that

$$B(\hat{\boldsymbol{\alpha}}) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\bar{\alpha})} \frac{\prod_{i=1}^{n} G_{a_i}\big(\prod_{j=0}^{a_i-1}(\alpha_i + j)\big)}{G_{\bar{a}}\big(\prod_{j=0}^{\bar{a}-1}(\bar{\alpha} + j)\big)}$$

$$= B(\boldsymbol{\alpha}) \frac{\prod_{i=1}^{n} G_{a_i}\big(\prod_{j=0}^{a_i-1}(\alpha_i + j)\big)}{G_{\bar{a}}\big(\prod_{j=0}^{\bar{a}-1}(\bar{\alpha} + j)\big)}. \tag{48}$$

Supplementary Material

Pairwise Non-Local Priors for CEG Model Selection: Supplementary Material (DOI: 10.1214/15-BA981SUPP; .pdf). The supplementary document includes the normalisation constants of pm-NLPs using Hellinger distance and its extension to ρ-norm space ($\rho \in \mathbb{N}+$), the computational results for all simulations presented here (Section 4) using the Hellinger pm-NLPs, and all CEG models found in Section 4.2 by the AHC algorithm using local and non-local priors.

References

Abramowitz, M. and Stegun, I. A. (1972). "Handbook of mathematical functions with formulas, graphs, and mathematical tables." National Bureau of Standards Applied Mathematics Series, 55, 10th printing (with corrections). MR0167642. 1193, 1197

Altomare, D., Consonni, G., and La Rocca, L. (2013). "Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors." *Biometrics*, 69(2): 478–487. MR3071066. doi: http://dx.doi.org/10.1111/biom.12018. 1166, 1180

Bangsø, O. and Wuillemin, P.-H. (2000). *Top-Down Construction and Repetitive Structures Representation in Bayesian Networks*, 282–286. AAAI Press. 1165

Barclay, L. M., Collazo, R. A., Smith, J. Q., Thwaites, P., and Nicholson, A. (2015). "The dynamic chain event graph." *Electronic Journal of Statistics*, 9(2): 2130–2169. MR3400535. doi: http://dx.doi.org/10.1214/15-EJS1068. 1193

Barclay, L. M., Hutton, J. L., and Smith, J. Q. (2013). "Refining a Bayesian Network using a Chain Event Graph." *International Journal of Approximate Reasoning*, 54(9): 1300–1309. MR3115418. doi: http://dx.doi.org/10.1016/j.ijar.2013.05.006. 1166, 1167, 1168, 1172

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. New York; Chichester: Wiley, 2nd edition. MR1700749. doi: http://dx.doi.org/10.1002/9780470316962. 1179, 1194, 1196

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). "Context-specific independence in Bayesian networks." In: Horvitz, E. and Jensen, F. (eds.), *12th Conference on Uncertainty in Artificial Intelligence (UAI'96)*, Uncertainty in Artificial Intelligence, 115–123. San Francisco: Morgan Kaufmann Publishers Inc. MR1617129. 1165

Bozga, M. and Maler, O. (1999). "On the Representation of Probabilities over Structured Domains." In: Halbwachs, N. and Peled, D. (eds.), *Computer Aided Verification*, volume 1633 of *Lecture Notes in Computer Science*, 261–273. Springer Berlin Heidelberg. MR1730246. doi: http://dx.doi.org/10.1007/3-540-48683-6_24. 1166

Collazo, R. A. and Smith, J. Q. (2015). Supplement to "A New Family of Non-Local Priors for Chain Event Graph Model Selection." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/15-BA981SUPP. 1182, 1184

Consonni, G., Forster, J. J., and La Rocca, L. (2013). "The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data." *Statistical Science* 28(3): 398–423. MR3135539. doi: http://dx.doi.org/10.1214/13-STS433. 1166

Consonni, G. and La Rocca, L. (2011). "On Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs." In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9 – Proceedings of the ninth Valencia international meeting*, 63–78. Oxford University Press. MR3204454.

doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.001.0001. 1166, 1180

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. New York; London: Springer. MR1697175. 1165

Cowell, R. G. and Smith, J. Q. (2014). "Causal discovery through MAP selection of stratified chain event graphs." *Electronic Journal of Statistics*, 8(1): 965–997. MR3263109. doi: http://dx.doi.org/10.1214/14-EJS917. 1166, 1167, 1168, 1172, 1178, 1184

Cussens, J. (2008). "Bayesian network learning by compiling to weighted MAX-SAT." In: *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9–12, 2008*, 105–112. 1193

Cuthbertson, I. M. (2004). "Prisons and the Education of Terrorists." *World Policy Journal*, 21(3): pp. 15–22. 1190, 1191

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer. MR2664452. 1195

Dawid, A. (2011). "Posterior Model Probabilities." In: Bandyopadhyay, P. S. and Forster, M. R. (eds.), *Philosophy of Statistics*, volume 7, 607–630. Amsterdam: North-Holland. 1166

Dawid, A. P. (1999). "The Trouble with Bayes Factors." Technical report, University College London. 1166

Fergusson, D. M., Horwood, L. J., and Shannon, F. T. (1986). "Social and Family Factors in Childhood Hospital Admission." *Journal of Epidemiology and Community Health*, 40(1): 50–58. 1167

Freeman, G. and Smith, J. Q. (2011). "Bayesian MAP model selection of chain event graphs." *Journal of Multivariate Analysis*, 102(7): 1152–1165. MR2805655. doi: http://dx.doi.org/10.1016/j.jmva.2011.03.008. 1166, 1168, 1169, 1171

Hannah, G., Clutterbuck, L., and Rubin, J. (2008). "Radicalization or Rehabilitation. Understanding the challenge of extremist and radicalized prisoners." Technical Report TR 571, RAND Corporation. 1190, 1191

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves." *Journal of the American Statistical Association*, 101(473): 18–29. MR2252430. doi: http://dx.doi.org/10.1198/016214505000000187. 1171

Heckerman, D. (1999). "Learning in Graphical Models." chapter A Tutorial on Learning with Bayesian Networks, 301–354. Cambridge, MA, USA: MIT Press. 1172, 1176

Jaeger, M. (2004). "Probabilistic decision graphs – Combining verification and AI techniques for probabilistic inference." *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 12: 19–42. MR2058945. doi: http://dx.doi.org/10.1142/S0218488504002564. 1166

Jaeger, M., Nielsen, J. D., and Silander, T. (2006). "Learning probabilistic decision graphs." *International Journal of Approximate Reasoning*, 42(1–2): 84–100. MR2221585. doi: http://dx.doi.org/10.1016/j.ijar.2005.10.006. 1166

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society Series B – Statistical Methodology*, 72: 143–170. MR2830762. doi: http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x. 1166

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings (vol. 107 pg. 649, 2012)." *Journal of the American Statistical Association*, 107(500): 1656–1656. MR3036423. 1177

Jordan, J. and Horsburgh, N. (2006). "Spain and Islamist Terrorism: Analysis of the Threat and Response 1995-2005." *Mediterranean Politics*, 11(2): 209–229. 1190, 1191

Koller, D. and Pfeffer, A. (1997). "Object-oriented Bayesian Networks." In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, UAI'97*, 302–313. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. MR1464314. doi: http://dx.doi.org/10.1016/S0004-3702(97)00023-4. 1165

Korb, K. B. and Nicholson, A. E. (2011). *Bayesian Artificial Intelligence*. Chapman and Hall/CRC Computer Science and Data Analysis Series. Boca Raton, FL: CRC Press, 2nd edition. MR3100449. 1165, 1172

Kruskal, J. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika*, 29(1): 1–27. MR0169712. 1196

McAllester, D., Collins, M., and Pereira, F. (2008). "Case-factor diagrams for structured probabilistic modeling." *Journal of Computer and System Sciences*, 74(1): 84–96. MR2364183. doi: http://dx.doi.org/10.1016/j.jcss.2007.04.015. 1165

Ministry of Justice (2013). "Annual tables – Offender management caseload statistics 2012 tables." Online; accessed 03-Nov-2014. URL: https://www.gov.uk/government/statistics/offender-management-statistics-quarterly--2 1190

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Harlow: Prentice Hall. 1165

Neumann, P. E. (2010). "Prisons and Terrorism: Radicalisation and De-radicalisation in 15 Countries." Technical report, International Centre for the Study of Radicalisation and Political Violence, London. 1190, 1191

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press. MR2548166. doi: http://dx.doi.org/10.1017/CBO9780511803161. 1166

Poole, D. and Zhang, N. L. W. (2003). "Exploiting contextual independence in probabilistic inference." *Journal of Artificial Intelligence Research*, 18: 263–313. MR1996409. 1165

Rao, C. R. (1995). "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance." *Questiio*, 19(1–3): 23–63. MR1376777. 1196

Rowe, R. (2014). "From jail to jihad? The threat of prison radicalisation." Online; published 12-May-2014, accessed 19-Jan-2015. URL: http://www.bbc.co.uk/news/uk-27357208 1190

Schervish, M. (1996). *Theory of Statistics*. Springer Series in Statistics. Springer New York. MR1354146. doi: http://dx.doi.org/10.1007/978-1-4612-4250-5. 1195

Scutari, M. (2013). "On the Prior and Posterior Distributions Used in Graphical Modelling." *Bayesian Analysis*, 8(3): 505–532. MR3102220. doi: http://dx.doi.org/10.1214/13-BA819. 1166

Silander, T. and Leong, T.-Y. (2013). "A Dynamic Programming Algorithm for Learning Chain Event Graphs." In Fürnkranz, J., Hüllermeier, E., and Higuchi, T. (eds.), *Discovery Science*, volume 8140 of *Lecture Notes in Computer Science*, 201–216. Springer Berlin Heidelberg. 1167, 1172

Silke, A. (2011). *The Psychology of Counter-Terrorism*. Cass Series on Political Violence. Abingdon, Oxon, England; New York: Routledge. 1190, 1191

Smith, J. Q. (2010). *Bayesian Decision Analysis: Principles and Practice*. Cambridge; New York: Cambridge University Press. MR2828346. doi: http://dx.doi.org/10.1017/CBO9780511779237. 1165, 1168

Smith, J. Q. and Anderson, P. E. (2008). "Conditional independence and chain event graphs." *Artificial Intelligence*, 172(1): 42–68. MR2388535. doi: http://dx.doi.org/10.1016/j.artint.2007.05.004. 1166, 1168

Thwaites, P. A., Smith, J. Q., and Cowell, R. G. (2008). "Propagation using Chain Event Graphs." In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, 546–553. Corvallis, Oregon: AUAI Press. 1166, 1168

Acknowledgments

The authors wish to thank David Rossell for his valuable comments and, John Horwood and the CHDS research group for providing one of the data sets used in this paper. The authors would like also to thank the reviewers and editors of the journal for their insightful comments which have greatly improved this paper. One author was supported by the Brazilian Navy and CNPq-Brazil [grant number 229058/2013-2].